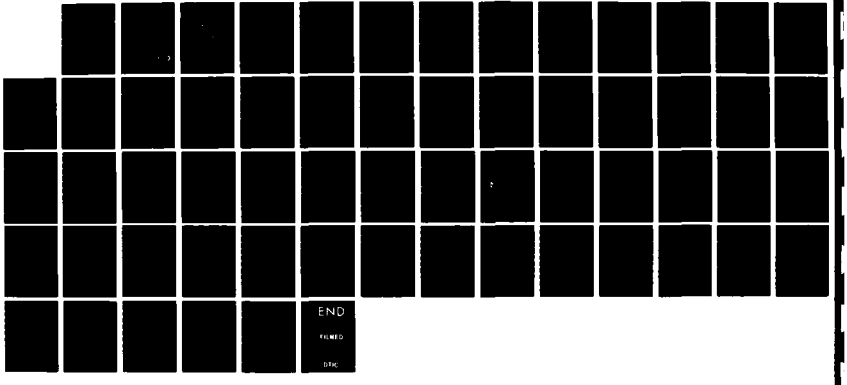
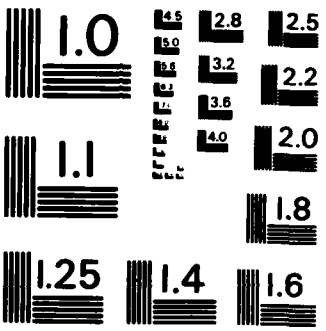
AD-A161 743 A COHERENT VLSI DESIGN ENVIRONMENT(U) MASSACHUSETTS 1/1
INST OF TECH CAMBRIDGE DEPT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE P PENFIELD ET AL. 30 SEP 85
UNCLASSIFIED N00014-80-C-0622 F/G 9/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MASSACHUSETTS 02139

A COHERENT VLSI DESIGN ENVIRONMENT

Semiannual Technical Report
for the period
April 1, 1985 to September 30, 1985

Massachusetts Institute of Technology
Cambridge, MA 02139

Principal Investigators:		
	Paul Penfield, Jr.	(617) 253-2506
	Lance A. Glasser	(617) 253-4677
	Thomas F. Knight, Jr.	(617) 253-7807
	Charles E. Leiserson	(617) 253-5833
	Ronald L. Rivest	(617) 253-5880
	John L. Wyatt, Jr.	(617) 253-6718
	Richard E. Zippel	(617) 253-6028

This research was sponsored by Defense Advanced Research Projects Agency (DoD), through the Office of Naval Research under Contract N00014-80-C-0622.

REPRODUCED AT GOVERNMENT EXPENSE

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

TABLE OF CONTENTS

Mark Douglas Matson, Macromodeling and Optimization of Digital MOS VLSI Circuits, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, January 24, 1985.*

Thang Nguyen Bui, Soma Chaudhuri, F. Thomson Leighton, and Michael Sipser, "Graph Bisection Algorithms with Good Average Case Behavior," MIT VLSI Memo No. 85-236, March 1985.*

Lance A. Glasser and Charles A. Zukowski, "Communication Density Considerations in Multiprocessors," MIT VLSI Memo No. 85-242, April 1985.

Lance A. Glasser, "A UV Write-Enabled PROM," Proceedings of the 1985 Chapel Hill Conference on VLSI (May 1985), Computer Science Press, pp. 61-65.

Zukowski, C., J. L. Wyatt, Jr., and L. A. Glasser, "Bounding Techniques and Applications for VLSI Circuit Simulation," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June 5-7, 1985, pp. 163-166.

Q. Yu, J. L. Wyatt, Jr., C. Zukowski, H. N. Tan and P. O'Brien, "Improved Bounds on Signal Delay in Linear RC Models for MOS Interconnect," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June 5-7, 1985, pp. 903-906.

G. C. Clark and R. E. Zippel, "Schema -- An Architecture for Knowledge Based CAD," to appear in Proc. IEEE International Conference on Computer-Aided Design, Santa Clara, CA, November 18-21, 1985.

J. L. Wyatt, Jr., C. A. Zukowski, and P. Penfield, Jr., "Step Response Bounds for Systems Described by M-Matrices, with Application to Timing Analysis of Digital MOS Circuits," to appear in Proceedings of the 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, December 1985.

---

* Abstract only. Complete version available from Microsystems Research Center, Room 39-321, MIT, Cambridge, MA 02139; telephone (617) 253-7308.

RESEARCH OVERVIEW

This report covers the period from April 1, 1985 through September 30, 1985. The research discussed here is described in more detail in several published and unpublished reports cited below.

The CAD frame Schema has progressed to the point where it is useful for ample chip designs. The interface to CIF is complete, and work has begun on importing layout libraries. An interface to EDIF is being installed. Simulators can now be connected, and thought is going into organization of VLSI libraries. A plan for the distribution of Schema is now being worked out. Members of the DARPA VLSI community will be able to get copies in the Fall of 1985 or Spring of 1986.

Previous results on waveform bounding have been generalized to large classes of problems described in canonical control-theory form. Work has begun on models for interconnect taking account of line inductance. This domain is less general that RLC networks, and there is hope that some of the previously derived bounds still apply. Indeed, some such results are reported here.

During this period a novel device, the UV write-enabled PROM, was reported at a conference. Work continues on developing useful circuits employing this device.

A new program in CAR (Computer-Aided Reliability) of VLSI circuits has begun. The initial emphasis is on modeling three effects, metal migration, time-dependent dielectric breakdown, and hot-electron effects. The objective is to create software which analyzes layouts for possible reliability problems and gives advice to the designer.

Work continues on the promising "fat-tree" interconnection network. Some improved message-delivery algorithms have been developed. This and a variety of improvements in routing, arithmetic and graph algorithms, and communications networks are described in this report.

THE DESIGN OF SCHEMA

Over the past six month a great deal of effort has gone into Schema -- enough that we have begun to use it for some sample chip designs. Our colleagues at Harris have made good progress on the PC board development system, and the simulation environment is beginning to be used.

Kerry O'Neill spent this summer developing routines for dealing with CIF for Schema. As a result Schema can now read and write its layout information in CIF. In addition, Kerry has begun developing a library of lambda based CMOS and NMOS layouts based on those contained in the Newkirk and Mathews book, "The VLSI Designer's Library." Texas Instruments has contributed an EDIF parser and reader for Schema which Kerry is now integrating into the rest of the system.

Margaret St. Pierre has been continuing her work on the Schema simulation environment. The waveform database has been completed and is being used in several ongoing projects. Margaret has just finished the Schema interface to Spice, where Schema makes use of a remote Spice server via the network. At the moment a VAX running Unix is being used but we are exploring the use of an IBM 4381 as a faster alternative. Chris Terman wrote a quick forward Euler simulator that has been incorporated. This simulator is best used as an illustration of interfacing with Schema, but is can also used for small, simple circuits by the emerging AI tools.

The PC board development sub-system raised an interesting problem which we would also have building VLSI libraries: How to organize modules in a fashion that would allow their selection by their behavior. For instance, a designer might be using 2-input NAND gates. When it comes time to bind the logic gate to a particular package, the designer might wish to know what TTL packages have 2-input NAND gates with less than 6 nanosecond propagation delays. Thus a simple hierarchical structure is not adequate for representing the relationships among the different package types. To deal with this problem we have being using the Kandor knowledge representation language developed at the Schlumberger Palo Alto Research Center as a knowledge indexed database into our hierarchy of Schema modules. Kandor databases are able to be components of a design hierarchy and thus provide an alternative indexing mechanism for modules in Schema. This structure seems to be working quite well, and we are now evaluating different strategies for fully integrating it with Schema. A more detailed report on this work will follow.

Schema's usefulness is steadily increasing, and the number of requests for copies has been escalating. We plan to make copies available to patient and indulgent colleagues this fall, and to make it more widely available in Spring 1986, after we have run a few designs through it.

THE WAVEFORM BOUNDING APPROACH TO DELAY ESTIMATION

We have recently achieved progress on several fronts. In the area of delay bounds for on-chip interconnect, we have finally proved a conjecture which, at a very modest cost in computer time, enables us to further tighten the original bounds[1] on signal delay in RC tree networks. We have discovered the theoretical foundations for a class of waveform bounding techniques and extended them to include a more general class of dynamic systems. We have initiated a project to extend the methods to bipolar ECL circuits and successfully completed the first part of it. And we have discovered several practical tricks to speed up computation.

We have begun studying the more difficult problem of delay bounds for inter-chip interconnect with significant inductance and have proved two enlightening preliminary results. We have also found and rigorously justified a promising method for "relaxing" bounds to produce tighter ones. More complete summaries are given in the individual sections below.

In the area of RC models for on-chip interconnect we have found a way to further tighten the original bounds[1] by exploiting the spatial convexity of interconnect voltage during transients in a novel way. A limited version of this idea was reported.[2] Mr. David Standley has discovered and proved a general version that is not yet published.

We had hoped to find that the original work[1] which stimulated this research does not rely in any fundamental way on many of the special properties of the RC tree networks to which it was applied. It turns out that its validity depends only on the sign pattern of the underlying system matrix, and thus it can be applied to a large class of dynamic systems in other areas of science and engineering.[3]

The research proposed for this grant was initially limited to MOS logic, but there is also an urgent need for fast timing analysis techniques suitable for ECL (emitter-coupled logic) bipolar chips. The project we have initiated to meet this need has three parts: i) extending the original work[1] to include networks with resistive leakage paths to ground modeling the base current

---

[1] J. Rubinstein, P. Penfield, Jr., and M. A. Horowitz, "Signal Delay in RC Tree Networks," IEEE Trans. CAD, vol. 2, no. 3, pp. 202-211, July, 1983.

[2] Q. Yu, J. L. Wyatt, Jr., C. Zukowski, H. N. Tan and P. O'Brien, "Improved Bounds on Signal Delay in Linear RC Models for MOS Interconnect," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June, 1985, pp. 903-906.

[3] J. L. Wyatt, Jr., C. A. Zukowski, and P. Penfield, Jr., "Step Response Bounds for Systems Described by M-Matrices, with Application to Timing Analysis of Digital MOS Circuits," to appear in Proceedings of the 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, December, 1985; also MIT VLSI Memo No. 85-257, September, 1985.

pathway in bipolar logic, ii) finding a rational and easily automated method for modeling the driving-point impedance of bipolar logic gates by simple RC circuits, and iii) finding a computationally fast way to calculate the required elements of the resistance matrix R for an interconnect network with leakage paths to ground. Mr. Peter O'Brien has completed the first of these three tasks this summer; a write-up will soon be available.

Advances to save computation time can be quite important in practice. One of these is based on the observation that the electrical models included with MOS standard-cell libraries commonly give two values for the cell's output resistance: one for output high-to-low transitions and another for low-to-high. The straightforward way of calculating the time constants needed in waveform bounding requires two entirely separate computations, one for each value of the driver resistance. But we have found a fast way of updating the initial computation to allow for a change in driver resistance that cuts this part of the computation time almost in half.

In the area of linear RLC models for inter-chip interconnect, these interconnect lines on printed circuit boards have much greater inductance than on-chip lines due to the absence of a nearby ground plane and, secondarily, their greater typical lengths. Fast methods of estimating and bounding propagation delay through such wires are urgently needed, but the techniques used in the early research[1] cannot be readily applied, primarily because the step response of lines with inductances exhibits overshoot and ringing. In searching for computationally fast methods to study delay in such lines, we have made some initial discoveries that are interesting in their own right and lead us to hope that more substantial progress will be possible. Mr. David Standley has found that for any such RLC line, possibly nonuniform and branched, the first moment of the impulse response (perhaps a reasonable estimate of delay in some cases) is utterly unaffected by the presence and numerical value of all inductors, and the second moment (a reasonable measure of the step response rise time in many cases) decreases monotonically with the value of each inductor.

In the area of bound relaxation, we have discovered a method by which the Waveform Relaxation techniques[4] developed for the efficient exact analysis of digital MOS circuits can be extended to include the relaxation of bounds.[5] In this project, bounding is considered as a framework within which to combine the latest algorithms for VLSI simulation in a form where uncertainty can be measured and therefore managed efficiently. The use of relaxation techniques allows partitioning to exploit latency and achieve a computation time that scales roughly linearly with circuit size. The use of simplified models to calculate response bounds allows efficient approximating algorithms, such as

---

[4] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, "The Waveform Relaxation Method for Time-Domain Analysis of Large Scale Integrated Circuits," IEEE Trans. CAD, vol. CAD-1, no. 3, pp. 131-145, July, 1982.

[5] Zukowski, C., J. L. Wyatt, Jr., and L. A. Glasser, "Bounding Techniques and Applications for VLSI Circuit Simulation," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June, 1985, pp. 163-166.

piecewise linear simulation, to be exploited. Bounding work for linear circuits can also be incorporated as one step in generating bounds on the responses of more complex circuits.

This last is a broadly focused and ambitious project. A number of exciting results have been obtained, but considerably more work is needed before it can serve as the basis for a practical CAD system.

HIGH PERFORMANCE CIRCUIT DESIGN

Four new write-enabled UV PROM designs have been submitted to MOSIS, three in nMOS and one in CMOS. One of the nMOS designs has an area of only 19 by 37 lambda. This design is similar to an nMOS static RAM. Another design, suggested by Prof. Paul Penfield, is similar to an LSSD latch. Should these designs prove to be functional and robust, they will be added to the MOSIS library.

The computer-aided circuit reliability effort is progressing nicely. The objective of this project is to build a circuit simulator that enables the designer to make choices among circuit configurations on the basis of reliability metrics such as the median time to failure. We now have preliminary models for the three phenomenon to be incorporated into the first simulator: metal migration, time-dependent dielectric breakdown, and hot-electron effects. A reliability simulator, RELIC, is planned to be built on top of an existing simulator. We are presently evaluating SAMSON from CMU and RELAX from Berkeley, for this purpose.

We have been developing a theory which predicts lower bounds on the conversion times of A/D converters as a function of the probability of a fault caused by synchronizer failures. We now have experimental verification of this phenomena. The general conclusion one reaches as a result of this study, is that flash converters do not have a significant advantage over self-timed successive approximation converters for extremely high reliability converters. This is because one spends the vast majority of the time waiting for the one slow bit (the one with the input near the metastable point of the comparator) to settle out.

With the help of Prof. John Wyatt, detail and rigor are being added to the work on reliability/noise margin/speed tradeoffs in digital MOS circuits.

ARCHITECTURAL DESIGN

Ron Greenberg and Professor Leiserson have continued to improve their algorithm for routing messages in the "fat-tree" interconnection network. The new algorithm follows the same general approach of randomizing in the choice of whether or not to send a particular message in a particular delivery cycle, but it achieves superior running times. As before, the performance of the algorithm is measured in terms of the load factor $\lambda(M)$ of the set M of messages to be routed, where the load factor can be defined in a general network setting as the maximum over all cuts in the network of the ratio of the number of messages which must pass through the cut divided by the capacity of the cut. For $\lambda(M) = \Omega(\lg n \lg\lg n)$, the new algorithm will, with high probability, route the set of messages, M, in $O(\lambda(M))$ delivery cycles, thereby meeting the lower bound. Various related results have also been obtained.

Alexander Ishii has developed a timing scheme for transmitting messages between processors in Professor Leiserson's "fat-tree" interconnection network. In addition, he and Bruce Maggs have implemented an NMOS VLSI fat-tree network interface utilizing the scheme. Work on fat-tree simulations has temporarily stopped, in anticipation of either new simulation facilities or a more optimized simulation program. Simulations with 8000 processors have already been run, however, and results so far appear encouraging. Currently, he and Professor Leiserson are investigating the verification and complexity of multiphase MOS clocking disciplines, in hopes of developing a theoretically sound basis for optimizing their use.

Cindy Phillips completed her master's thesis this summer. The thesis contains previously completed joint work with Professor Leiserson on finding the connected components of rectangles in the plane. It also presents a data structure that can be used in scanning-based algorithms which must maintain sets of segments. This data structure is much simpler than previous structures for this application. The data structure requires $O(n)$ space for n segments. Insertions, deletions, and finding one segment that overlaps a test segment can all be performed in $O(\lg n)$ time.

Miller Maley continued work on the mathematical foundations of the VLSI wiring model he developed with Professor Leiserson. This work aims to extend the applicability of Maley and Leiserson's new wire routing and layout compaction algorithms.

Tom Cormen and Professor Leiserson have designed a hyperconcentrator switch for routing bit-serial messages in highly parallel routing networks. The switch, which has a highly regular layout, has been implemented in nMOS. It takes advantage of the relatively fast performance of large fan-in NOR gates in MOS technologies. The same architecture works for domino CMOS as well. A signal incurs exactly $2 \lg n$ gate delays through an n-input hyperconcentrator switch. This switch acts as a perfect concentrator switch and can be used to increase the performance of many existing routing networks.

Johan Hastad and Professor Leighton have developed improved bounds on the size of the smallest known circuits for division with O(log N) depth. Previously, the best circuits known for division of N-bit numbers in O(log N) steps had size $\Theta(N^5)$. The circuit developed by Hastad and Leighton also works in O(log N) steps but needs only $O(N^{1+\epsilon})$ gates, a significant improvement. The circuit is not yet practical, but the key ideas used to decrease the size of the circuit could well find application in other problems.

Bonnie Berger and Professor Leighton have developed an improved algorithm for 2-layer channel routing. For two-point nets, the algorithm uses $d+O(d^{1/2})$ tracks to route a problem with density d. For multipoint net problems with density d, the bound is $2d+O(d^{1/2})$. These bounds are a factor of two better than the best previously known bounds of Rivest, Baratz and Miller. The new algorithm allows wires to overlap for unit segments in the vertical direction (i.e., it uses the unit-vertical-overlap model). Most channel routing problems with density d require d tracks, even if arbitrary overlap is allowed, so the new results are very close to optimal in a strong sense.

Thang Bui and Professor Leighton continue to work on the graph bisection problem. They have now shown that greedy algorithms perform well (i.e., almost always find the optimal bisection) on random d-regular graphs having small bisection width, for large enough degree d. The work indicates that heuristics which use 2-change operations (such as the well known Kernighan-Lin heuristic) will also perform well for this class of graphs. It is hoped that this result can be extended to classes of graphs with larger bisection widths and smaller degrees. Already the work has led to improved heuristics for the graph bisection problem which appear to work very well on a wide variety of randomly generated graphs.

Professor Goldwasser is continuing her work on pseudo-random number generation. She, together with Dr. Goldreich and Professor Micali, showed how to construct pseudo-random functions from k bits to k bits which are indistinguishable from truly random functions in probabilistic polynomial (in k) time, based on the assumption that one-way functions exist. Pseudo-random functions can be used for generating test data for VLSI circuits.

Professor Goldwasser, Professor Micali, Benny Chor (who has just received his PhD) and Professor Awerbuch designed a method by which an n-node semi-synchronous broadcast network with o(log n) faults can be transformed into an n-node simultaneous broadcast network with o(log n) faults. It is especially easy to design fault-tolerant algorithms for the latter model.

PUBLICATIONS LIST

Mark Douglas Matson, Macromodeling and Optimization of Digital MOS VLSI Circuits, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, January 24, 1985; also MIT VLSI Memo No. 85-231, February 1985.

Thang Nguyen Bui, Soma Chaudhuri, F. Thomson Leighton, and Michael Sipser, "Graph Bisection Algorithms with Good Average Case Behavior," MIT VLSI Memo No. 85-236, March 1985.

T. Leighton, "Tight Bounds on the Complexity of Parallel Sorting," IEEE Trans. on Computers, vol. C-34, no. 4, April 1985, pp. 71-80.

F. Berman, T. Leighton, P. Shor and L. Snyder, "Generalized Planar Matching," MIT-LCS TM #273, April 1985.

Lance A. Glasser and Charles A. Zukowski, "Communication Density Considerations in Multiprocessors," MIT VLSI Memo No. 85-242, April 1985.

Lance A. Glasser, "A UV Write-Enabled PROM," Proceedings of the 1985 Chapel Hill Conference on VLSI (May 1985), Computer Science Press, pp. 61-65; also MIT VLSI Memo No. 85-239, March 1985.

F. M. Maley, "Compaction with Automatic Jog Introduction," Proceedings of the 1985 Chapel Hill Conference on VLSI (May 1985), Computer Science Press, pp. 261-283.

Lance A. Glasser, "Delay, Noise Margin, and Reliability in Digital Circuits, Proceedings of the 1985 Chapel Hill Conference on VLSI (May 1985), Computer Science Press, pp. 309-328; also MIT VLSI Memo No. 85-241, March 1985.

T. Leighton and C. Leiserson, "Wafer-Scale Integration of Systolic Arrays," IEEE Trans. on Computers, vol. C-34, no. 5, May 1985, pp. 448-461.

Zukowski, C., J. L. Wyatt, Jr., and L. A. Glasser, "Bounding Techniques and Applications for VLSI Circuit Simulation," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June 5-7, 1985, pp. 163-166; also MIT VLSI Memo No. 85-238, March 1985.

Q. Yu, J. L. Wyatt, Jr., C. Zukowski, H. N. Tan and P. O'Brien, "Improved Bounds on Signal Delay in Linear RC Models for MOS Interconnect," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June 5-7, 1985, pp. 903-906.

C. Phillips, "Space-Efficient Algorithms for Computational Geometry," M.S. Thesis, Department of Electrical Engineering and Computer Science, MIT, August 1985.

G. C. Clark and R. E. Zippel, "Schema -- An Architecture for Knowledge Based CAD," to appear in Proc. IEEE International Conference on Computer-Aided Design, Santa Clara, CA, November 18-21, 1985.

J. L. Wyatt, Jr., C. A. Zukowski, and P. Penfield, Jr., "Step Response Bounds for Systems Described by M-Matrices, with Application to Timing Analysis of Digital MOS Circuits," to appear in Proceedings of the 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, December 1985; also MIT VLSI Memo No. 85-257, September 1985.

Lectures Without Proceedings

Tom Leighton, "Random Matching Problems and the Average Case Analysis of Algorithms," Harvard University, April 22, 1985; MIT, April 29, 1985; University of Chicago Workshop on Computational Complexity Theory, May 2, 1985; and New York Academy of Sciences, June 13, 1985.

Paul D. Basset and Lance A. Glasser, "A High-Speed Asynchronous Communication Technique for MOS VLSI Systems," MIT VLSI Research Review, May 20, 1985.

Andrew A. Berlin, "A Digital Convolver for Visual Images," MIT VLSI Research Review, May 20, 1985.

Tom Cormen and Charles E. Leiserson, "A Hyperconcentrator Switch for Routing Bit-Serial Messages", MIT VLSI Research Review, May 20, 1985.

Ronald I. Greenberg and Charles E. Leiserson, "A Randomized Algorithm for Routing Bit-Serial Messages on Fat-Trees," MIT VLSI Research Review May 20, 1985.

Charles A. Zukowski, Lance A. Glasser, and John L. Wyatt, Jr., "Bounding Enhancements for VLSI Circuit Simulation," MIT VLSI Research Review, May 20, 1985.

Tom Leighton, "Tight Bounds of the Complexity of Parallel Sorting," Bell Communications Research - Morristown, August 8, 1985.

Tom Leighton, "Wafer-Scale Integration and the Grid Reconstruction Problem," Theory Day at Columbia University, September 27, 1985.

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-231 February 1985

Macromodeling and Optimization of Digital MOS VLSI Circuits*

Mark Douglas Matson**

ABSTRACT

Power consumption and signal delay are crucial to the design of high-performance VLSI circuits. This thesis presents CAD tools for modeling and optimizing digital MOS designs. The tools determine the transistor sizes that minimize circuit power consumption subject to constraints on signal path delays. Computational efficiency is obtained through macromodeling techniques and a specialized optimization algorithm. The macromodels are based on device equations, and encapsulate logic gate behavior in a set of simple yet accurate formulas. The optimization algorithm exploits properties of the digital MOS domain to convert the primal optimization problem into a dual form which is much easier to solve. The result is a pair of CAD tools that can optimize a circuit in roughly the amount of time needed to perform a transistor level simulation of the circuit.

---

*Submitted to the Department of Electrical Engineering and Computer Science on January 24, 1985 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy. This work was supported in part by an RCA fellowship, in part by the Defense Advanced Research Projects Agency of the Department of Defense under Contract No. N00014-80-C-0622, and in part by the Air Force Office of Sponsored Research under Contract No. F49620-84-C-0004.

**Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA 02139, Room 36-575, (617) 253-8169.

Macromodeling and Optimization of Digital MOS VLSI Circuits

by

Mark Douglas Matson

Submitted to the
Department of Electrical Engineering and Computer Science
on January 24, 1985 in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy.

Abstract

Power consumption and signal delay are crucial to the design of high-performance VLSI circuits. This thesis presents CAD tools for modeling and optimizing digital MOS designs. The tools determine the transistor sizes that minimize circuit power consumption subject to constraints on signal path delays. Computational efficiency is obtained through macromodeling techniques and a specialized optimization algorithm. The macromodels are based on device equations, and encapsulate logic gate behavior in a set of simple yet accurate formulas. The optimization algorithm exploits properties of the digital MOS domain to convert the primal optimization problem into a dual form which is much easier to solve. The result is a pair of CAD tools that can optimize a circuit in roughly the amount of time needed to perform a transistor level simulation of the circuit.

Thesis Supervisor: Lance A. Glasser

Title: Assistant Professor of Electrical Engineering and Computer Science

Table of Contents

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-236 March 1985

Graph Bisection Algorithms With Good Average Case Behavior*

Thang Nguyen Bui, Soma Chaudhuri, F. Thomson Leighton, Michael Sipser**

ABSTRACT

In the paper, we describe a polynomial time algorithm that, for every input graph, either outputs the minimum bisection of the graph or halts without output. More importantly, we show that the algorithm chooses the former course with high probability for many natural classes of graphs. In particular, for every fixed $d \geq 3$, all sufficiently large n and all

$$b = o(n^{1-1/\mathrm{floor}((d+1)/2)}),$$

the algorithm finds the minimum bisection for almost all d-regular labelled simple graphs with 2n nodes and bisection width b. For example, the algorithm succeeds for almost all 5-regular graphs with 2n nodes and bisection width $o(n^{2/3})$. The algorithm differs from other graph bisection heuristics (as well as from many heuristics for other NP-complete problems) in several respects. Most notably:

(i) the algorithm provides exactly the minimum bisection for almost all input graphs with the specified form, instead of only an approximation of the minimum bisection,
(ii) whenever the algorithm produces a bisection, it is guaranteed to be optimal (i.e., the algorithm also produces a proof that the bisection it outputs is an optimal bisection),
(iii) the algorithm works well both theoretically and experimentally,
(iv) the algorithm employs global methods such as network flow instead of local operations such as 2-changes, and
(v) the algorithm works well for graphs with small bisections (as opposed to graphs with large bisections, for which arbitrary bisections are nearly optimal).

---

*This research was supported by Air Force Contract No. AFOSR-82-0326 and DARPA Contract No. N00014-80-C-0622.

**Mathematics Department and Laboratory for Computer Science, MIT, Cambridge, MA 02139; Bui: Room NE43-838, (617) 253-5971; Leighton: Room 2-372, (617) 253-3662; Sipser: Room 2-376, (617) 253-4989; Chaudhuri: Department of Computer Science, University of Washington, Seattle, Washington.

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-242 April 1985

Communication Density Considerations in Multiprocessors*

Lance A. Glasser and Charles A. Zukowski**

ABSTRACT

Fundamental limits on the communication capabilities of large multiprocessors is investigated. It is shown that, in large homogeneous machines with isotropic communication, interprocessor communication must fall off faster than the fourth power of distance. The effects of scaling on information density is also considered.

---

*This research was supported in part by the National Science Foundation under grant number ECS-8118160, by the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622, and by an IBM Graduate Fellowship.

**Department of Electrical Engineering and Computer Science and Research Laboratory of Electronics, MIT, Cambridge, MA 02139, Glasser: Room 36-587, (617) 253-4677, Zukowski: Room 36-585, (617) 253-8169.

Communication Density Considerations in Multiprocessors

by

Lance A. Glasser
and
Charles A. Zukowski

Electrical Engineering and Computer Science Department
and the
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Abstract: Fundamental limits on the communication capabilities of large multiprocessors is investigated. It is shown that, in large homogeneous machines with isotropic communication, interprocessor communication must fall off faster than the fourth power of distance. The effects of scaling on information density is also considered.

Key words: Multiprocessors, Communication limits, Information density, Massive parallelism

---

* This research was supported in part by the National Science Foundation under grant number ECS-8118160, by the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622, and by an IBM Graduate Fellowship.

Communication Density Considerations in Multiprocessors

by
Lance A. Glasser
and
Charles A. Zukowski

1 Introduction

Massively parallel fine-grained multiprocessors [1,2] require massive amounts of communication. Interprocessor communication bandwidth is thought to strongly impact the performance of multiprocessors by imposing limits on the degree to which processors can exchange information and cooperate on a single problem. In this paper we examine the physical limits imposed on interprocessor communication by the capacity of the communication medium. Other constraints, such as those discussed in [3], provide additional limitation on the performance of large digital multiprocessors. In the spirit of massive parallelism, we use continuous rather than discrete variables in the formulation of the model. This choice makes the mathematics cleaner without sacrificing the physics.

We know, from information theory, that the transmission of one bit of information requires on the order of energy kT, where k is Boltzman's constant and T is absolute temperature. Thus, high data rates require large quantities of power, independent of how the information is transmitted or coded. If we examine a transmission medium of fixed diameter, whether it be a length of coaxial cable or a fiber optic link, the information velocity is limited to c, the speed of light. If the "wire" diameter is D, then the energy density in the wire must be at least $BE_0/(cD^2)$, where B is the bit rate and E_0 is the minimum energy required to transmit one bit. All physical media, with the exception of absolute vacuum, have a maximum energy density that they can withstand without breaking down. Since electromagnetic energy is usually the information carrying medium of choice, limits on the energy density often arise out of electric forces which become significant when compared to the forces binding electrons to their atoms. These electric fields will cause the medium to behave non-linearly, distorting the information. Large enough fields will cause destruction of the medium. One can counteract this phenomena by either making D larger or running several cables in parallel. We therefore argue that a fundamental quantity which is physically limited is information density. The highest values of information density are found in optical fibers and VLSI chips, where values of $10^{19\pm2}\mathrm{b}/(\mathrm{s}\cdot\mathrm{m}^2)$ can be observed. Usually communication densities are much lower.

While information densities may be far from their fundamental limits is current technologies, there will always be practical limits and costs associated with different densities. Therefore, it is important to study the general relationship between information densities and interprocessor communication. Given an array of processors and a communication medium of bounded volume, the interprocessor communication requirements necessarily

generate information density requirements on the medium. Limits on information density therefore impose limits on the degree to which processors can communicate and hence cooperate. It is these limits which we will discuss in this paper.

2 Statement of Problem

Assume that K dimensional space is filled with processing elements with a density ρ, where ρ is a function of position. Processors in two dimensional space might represent an array on a planar integrated circuit. Large systems, consisting of many VLSI sub-systems, are constructed to take advantage of all three physical dimensions. In this section, we present the machinery necessary to analyze interprocessor communication in such an array. For simplicity we assume that all functions discussed are continuous and bounded.

First consider the information flowing out from a single processing element. The communication requirement for the entire array is the sum over all elements of these outgoing information flows. If all communication between every two processors is assumed to be through the shortest path, the flux density of bandwidth from a single processor is a radial vector field. That is, each processor emits, along a radial line, that information required by other processors in that direction. Near the transmitter, the information density due to the transmitter must contain the information intended for all receivers in a given direction. Further away from the transmitter, the information flux decreases as various receivers remove data. $\vec{F}(\vec{q},\vec{s})$ represents the information flux density at point $\vec{q}$ originating from a processor at point $\vec{s}$. $f(\vec{q},\vec{s})$ denotes the scalar magnitude of the flux and has the units $\mathrm{b/(s{\cdot}proc{\cdot}m^{K-1})}$. We have the simple relation

$$\vec{F}(\vec{q},\vec{s}) = \frac{(\vec{q}-\vec{s})}{|\vec{q}-\vec{s}|} f(\vec{q},\vec{s}). \tag{1}$$

The flux from a single processor can be viewed as a vector field, but the contributions from different processors do not add as vectors. If two processors are exchanging information at equal rates, the total communication taking place is not zero. As a result, our definition assumes that the flux magnitude is always positive and incoming information is represented in the positive outgoing flux from other processors. The function $f(\vec{q},\vec{s})$ must be a decreasing function of the distance $|\vec{q}-\vec{s}|$ because other processors can only receive information from the processor at point $\vec{s}$. In any dimension larger than one, the surface area available for communication increases with distance. Therefore, when non-zero, f must fall with distance at least as fast as $|\vec{q}-\vec{s}|^{1-K}$, the rate seen in the absence of receiving processors. Figure 1 illustrates the relevant geometries. In the absence of receivers, the total information travelling outward from $\vec{s}$ is independent of distance, but the information density is not.

To express the notion of receiving processors we define the quantity $I(\vec{q},\vec{s})$ to represent the information bandwidth flowing from a processor at point $\vec{s}$ to one at point $\vec{q}$. I has the units of $\mathrm{b/(s{\cdot}proc^2)}$. Intuitively, the information flowing to a processor at $\vec{q}$ is the amount of flux disappearing there, scaled by the processor density at $\vec{q}$. This can be expressed as

$$\rho(\vec{q})I(\vec{q},\vec{s}) = -\nabla\cdot\vec{F}. \tag{2}$$

Using polar coordinates defined about the point $\vec{s}$, (2) can be re-expressed in integral form as

$$f(\vec{q},\vec{s}) = \frac{1}{|\vec{q}-\vec{s}|^{K-1}} \int_{|\vec{q}-\vec{s}|}^{\infty} \alpha^{K-1} \rho\left(\vec{s} + \alpha \frac{\vec{q}-\vec{s}}{|\vec{q}-\vec{s}|}\right) I\left(\vec{s} + \alpha \frac{\vec{q}-\vec{s}}{|\vec{q}-\vec{s}|}, \vec{s}\right) \mathrm{d}\alpha. \tag{3}$$

This form assumes that the boundary condition of zero information flux at infinity is satisfied, that is, all information sent by a processor is received by others.

To find the total communication density at a particular point, we add the magnitudes of the information fluxes originating from all processors. The total information flux through any point $\vec{q}$ is given by

$$\Phi(\vec{q}) = \int_V f(\vec{q},\vec{s})\rho(\vec{s}). \tag{4}$$

Φ represents the density of bits per second passing through an infinitesimal area of space and has the units of bits/(s·m^{K-1}). Said another way, Φ represents information rate density. It is this density which, we argued in the introduction, is a physically significant and limited quantity. Equation (4) can be simplified by assuming spherical symmetry about $\vec{q} = 0$ and considering only the bandwidth density at the origin. As a result of symmetry, the only necessary coordinate is r, the distance from the origin. Equation (4) then becomes

$$\Phi(0) = C_K \int_0^{\infty} \left(r^{K-1}\right) f(0,r)\rho(r)\mathrm{d}r, \tag{5}$$

where

$$C_K = \begin{cases} 2 & \text{for } K = 1 \\ 2\pi & \text{for } K = 2 \\ 4\pi & \text{for } K = 3 \\ \dfrac{(K-3)C_{K-1}C_{K-2}}{(K-2)C_{K-3}} & \text{in general.} \end{cases} \tag{5.1}$$

Combining (3) and (5), we obtain

$$\Phi(0) = C_K \int_0^{\infty} \rho(r) \int_r^{\infty} (r^{*K-1})\rho(r^* - r) I(r - r^*, r)\mathrm{d}r^*\mathrm{d}r. \tag{6}$$

These geometries are illustrated in Fig. 2. Due to spherical symmetry, $I(-b,a)$ refers to the communication from a processor at a distance a from the origin to one a distance b from the origin on the opposite side. Also by the symmetry assumption, $\rho(r) = \rho(-r)$.

Equation (6) is the central result of the paper. In the remaining sections we consider some important special cases. First we look at the limit to global communication imposed by a finite communication density constraint. Then we look at how communication density scales with processor and communication distribution.

3 Finite Communication Density Constraint

The first special case of interest is that of ρ constant and $I(\vec{q},\vec{s})$ a function only of the euclidean distance $d = |\vec{q} - \vec{s}|$. That is, we have a model in which all of space is filled with processors, each communicating with all of the others. We assume, however, that the emphasis is on local (d small) communication. Our objective is to quantify the requirement for locality. Assume that $\rho = \rho_0$ and that $I(d)$ is a continuous bounded function of non-negative d such that

$$I(d) < \gamma(d_0 + d)^{-M} \tag{7}$$

for some positive constants γ, d_0, and M. In this case we say that communication falls with order M. We have

$$\Phi(0) = C_K \rho_0^2 \int_0^\infty \int_r^\infty r^{*K-1} I(r^*) dr^* dr, \tag{8}$$

and therefore

$$\Phi(0) < \gamma C_K \rho_0^2 \int_0^\infty \int_r^\infty (d_0 + r^*)^{K-1-M} dr^* dr. \tag{9}$$

As a result of (9), $\Phi(0)$ will converge if we meet the constraint that

$$M > K + 1. \tag{10}$$

Specifically, in two dimensions the communication must fall off with order greater than 3 and in three dimensions, the order must exceed 4. d^{-4} is a very rapidly decaying function and indicates the extreme penalty for long distance communication. This constraint is so severe because each processor on one side of the origin must communicate with every processor on the other side of the origin. Equation (6) limits the degree to which even an infinite number of processors can cooperate on a single problem.

The second special case of interest is that of I constant and ρ some continuous bounded function of the distance from the origin r. Many multiprocessors, such as the BBN Butterfly, strive to keep I constant so that the programmer does not have to deal with the added complexity of communication locality. Assume that $I = I_0$. We have

$$\Phi(0) = C_K I_0 \int_0^\infty \rho(r) \int_r^\infty r^{*K-1} \rho(r^* - r) dr^* dr, \tag{11}$$

and therefore

$$\Phi(0) > C_K I_0 \int_0^\infty \rho(r) \int_0^\infty r^{K-1} \rho(r^*) dr^* dr. \tag{12}$$

Noting that the total number of processors N is given by

$$N \equiv \int_V \rho(\vec{s}) = C_K \int_0^\infty r^{K-1} \rho(r) dr, \tag{13}$$

we see that

$$\Phi(0) > I_0 N \int_0^\infty \rho(r) dr. \tag{14}$$

From this it can be concluded that, with uniform communication, a finite communication density requires a finite number of processors, as one would expect. For N to be finite, $\rho(r)$ must fall off with an order larger than K.

4 Communication Scaling

Assuming that a processor array is finite, equation (6) also indicates how communication density scales with bandwidth requirements and array size. Consider the special case where ρ is constant over a finite array of radius R. Assume as before that I is only a function of distance d. Let $I(d)$ have the constant value $I_0 a^{-M}$ for $0 < d \leq a$ and be $I_0 d^{-M}$ for $d > a$, where a is some constant such that $0 < a < R$. From (6) we arrive at

$$\Phi(0) = C_K \rho_0^2 \int_0^R \int_r^{r+R} (r^{*K-1}) I(r^*) \mathrm{d}r^* \mathrm{d}r. \tag{15}$$

Evaluating (15), we find that

$$\Phi(0) = C_K I_0 \rho_0^2 \left(\beta + \delta R^{K-M+1}\right) \tag{16}$$

where

$$\beta = \frac{M a^{K-M+1}}{(K+1)(M-K-1)}, \tag{16.1}$$

$$\delta = \frac{2(2^{K-M}-1)}{(M-K)(M-K-1)}, \tag{16.2}$$

when M is not equal to K or $K+1$. Consistent with the results of the previous section, if communication falls off with an order larger than $K+1$, $\Phi(0)$ approaches asymptotically to the finite value of $C_K I_0 \rho_0^2 \beta$ as R is increased towards infinity. If M is smaller than $K+1$, (16) determines the rate at which $\Phi(0)$ approaches infinity with increasing radius R.

In the special case where $M = 0$ (I a constant function of distance), we find that $\beta = 0$. If the number of processors N is held constant, we have

$$\Phi(0) = \left(\frac{2K(2^K-1)}{C_K(K+1)}\right) N^2 I_0 R^{1-K} \tag{17}$$

or

$$\Phi(0) \propto N^{\frac{K-1}{K}} \rho^{\frac{K-1}{K}}. \tag{18}$$

From (18) we can see how communication density decreases with spreading of a processor array. With uniform communication among a fixed number of processors, the communication density falls off as the volume to the power $(1-K)/K$. For a two dimensional array on a planar integrated circuit, every factor of four increase in the area produces a factor of two reduction in the communication density. If, on the other hand, the processor density is held constant and the number of processors is increased by a factor of four, the communication density is increased by a factor of eight.

5 Conclusion

We have presented a continuous model for communication density in large multiprocessor arrays. Physical systems, regardless of their architecture, must contend with the limits we described. While our analysis is only valid for straight line communication, the average value of Φ can only increase when line-of-sight communication is not used.

Analysis of the model demonstrates the high costs of communication. This does not mean that global communication is completely unwarranted. As pointed out by Leiserson [4], there are some algorithms for which a little global communication is worth a lot of local communication. Similarly, we may also trade off greater numbers of processors with local communication against fewer numbers of processors with more global communication bandwidth. In the context of any one problem, the number of processors, together with the spatial characteristics and capacity for information exchange, can be evaluated under the integral in (6) to determine the communication density costs of a particular implementation.

References

[1] W. D. Hillis, "The Connection Machine," TR. 646, Cambridge, Mass., M.I.T. AI Laboratory, 1981.

[2] S. E. Fahlman, G. E. Hinton, and T. J. Sejnowki, "Massively Parallel Architectures for AI: Netl, Thistle, and Boltzmann Machines," *Proc. National Conf. Artificial Intelligence*, pp. 109–113, Washington D.C., 1983.

[3] R. W. Keyes, "Fundamental Limits in Digital Information Processing," *Proc. IEEE*, vol. 69, pp. 267–278, 1981.

[4] Leiserson, C. E., private communication.

Figure Captions

Fig. 1 The information flux leaving a point $\vec{s}$ is a decreasing function of distance.
Fig. 2 $\Phi(0)$ is the sum of the communications from all processors $\vec{s}$ to those processors $\vec{q}$ on the other side of the origin.

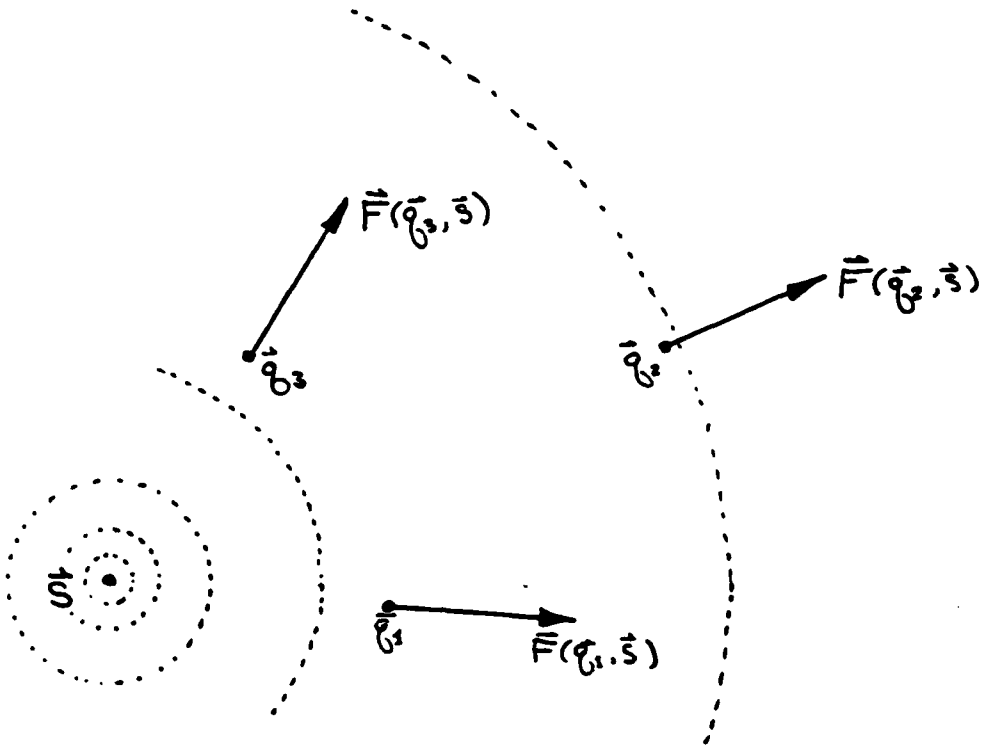

$\vec{F}(\vec{q}_3, \vec{S})$
$\vec{F}(\vec{q}_2, \vec{S})$
$\vec{q}_3$
$\vec{q}_2$
$\vec{S}$
$\vec{q}_1$
$\vec{F}(\vec{q}_1, \vec{S})$

FIG. 1

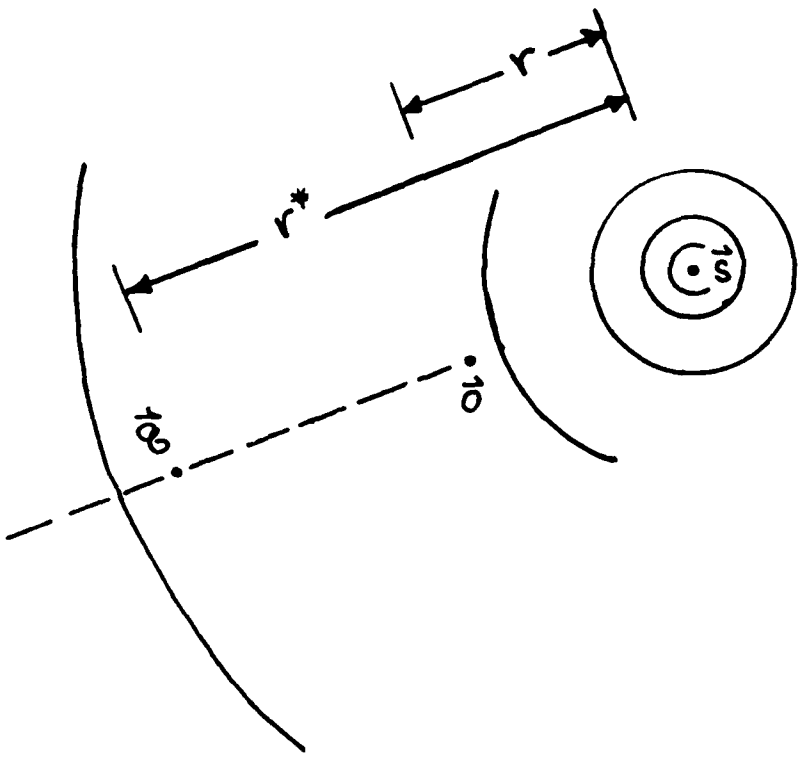


Fig. 2

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-239 March 1985

A UV Write-Enabled PROM*

Lance A. Glasser**

ABSTRACT

We have demonstrated a programmable read only memory cell that is written in the presence of unfocused UV light. Once the UV light is removed, the programmed state is non-volatile. The cell uses conventional MOS processing.

---

*To Appear in Proceedings, 1985 Chapel Hill Conference on VLSI, University of North Carolina at Chapel Hill, May 15-17, 1985. This research was supported by the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622.

**Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, MIT, Room 36-587, Cambridge, MA 02139, (617) 253-4677.

A UV Write-Enabled PROM

Lance A. Glasser*
Department of Electrical Engineering and Computer Science
and the
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Abstract: *We have demonstrated a programmable read only memory cell that is written in the presence of unfocused UV light. Once the UV light is removed, the programmed state is non-volatile. The cell uses conventional MOS processing.*

We have demonstrated a programmable read only memory cell that can be written into the "1," "0," or "previous" state in the presence of unfocused UV light. The programmed state is controlled by low electrical voltages. Once the UV light is removed, the programmed state is non-volatile. The memory cell uses conventional MOS processing with no additional mask steps. The cell can thus be implemented on virtually all silicon gate nMOS and CMOS processes. It was demonstrated in 4 μm nMOS using silicon foundry resources through the MOSIS facility. Functional chips were fabricated by Comdial and AMI.

The cell, illustrated schematically in Fig. 1, takes advantage of the fact that under UV excitation electrons can surmount the potential barrier at the silicon/silicon dioxide interface. Therefore, when the light impinges on the gate/source or gate/drain region of a MOS transistor, photo-excited electrons flow through the gate oxide so as to equalize the quasi-Fermi levels. We write both binary states rather than simply erase (write a "0" into) a cell, as had been done in previous UV-PROMs. In the circuit in Fig. 1, node v_{bit} is implemented entirely in polysilicon. This

* This research was supported by the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622.

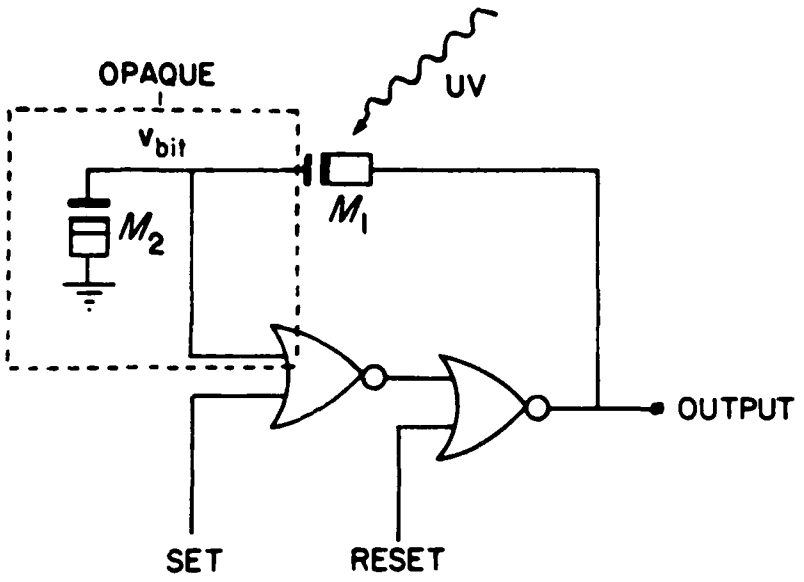


Fig. 1 Schematic of UV Write-enabled PROM.

is the floating gate. Only the floating gate area over the MOS capacitor M_1 is open to illumination. The rest of the node needs to be covered by opaque material; in most implementations this would be aluminum. A realistic layout is shown in Fig. 2. One must be careful not only to cover M_2 and M_3 with metal, but also one must cover the poly over field region.

All reading and writing is accomplished with power supplied to the chip. With nMOS technology this presents no problem unless one is counting on dynamic charge storage during the write pulse. With CMOS technology one must beware of photon induced latch-up; however, this usually requires much higher light intensities than we are using. Write times as short as of 10 min. were observed. During this time all cells may be changed simultaneously, if desired.

To write a "1," one raises the SET line and floods the chip with UV light. v_{bit} then charges up to a high voltage, ideally V_{DD}, but typically about 1.5 V lower. V_{DD} is the power supply voltage.

Experiments with 254 nm and 302 nm light indicate that the electrons that charge and discharge the gate are not coming from the meagerly populated silicon conduction bands, but rather from the valence bands where electrons are much more numerous. The conduction band to oxide barrier is usually taken as 3.1 eV. If the carriers were coming from the conduction band then both wavelengths, corresponding to 4.8 eV and 4.1 eV, would enable write action in the PROM. We observed only the shorter wavelength light to be effective. This is consistent with assum-

ing the carriers come from the valence band, in which case the barrier is about 4.2 eV—too high for electrons excited with the longer wavelength light.

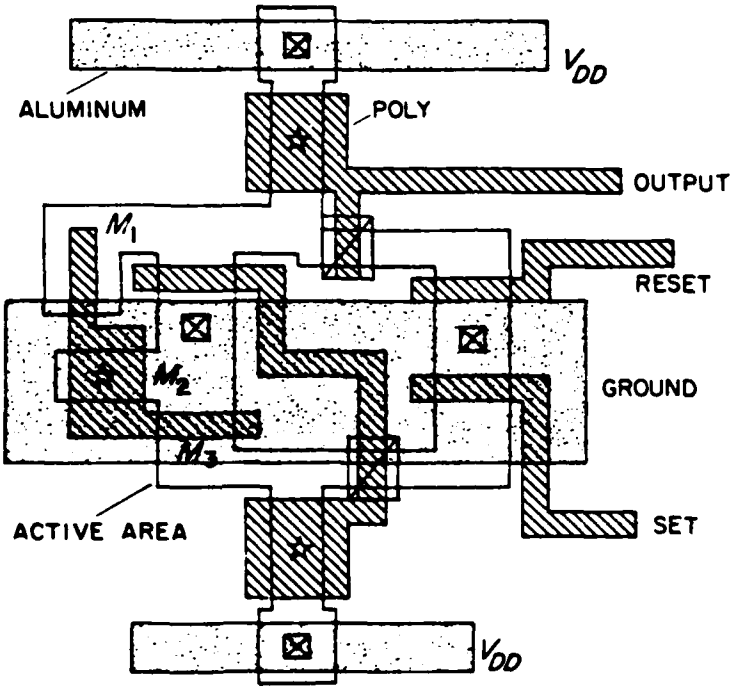


Fig. 2 Layout of PROM cell.

To write a "0" one raises the RESET line instead of the SET line. When the UV excitation is removed, the data is stored. To write the previous state, and this is important is cases when one only wants to change a few bits of the programming, one leaves the SET and RESET lines low while power and illumination are applied. The two cascaded NOR gates will assure that the last stored value is retained.

Many different cells are possible. The simplest (and most compact) cells do not have the ability to hold the last state while the other cells are being written; they can only be set or reset. A cell, such as the one in Fig. 2, can write the last state, but the SET or RESET lines must be held valid for a significant part of the write cycle.

To speed the writing of large arrays, one can write the data first into shadow registers. Cells with shadow registers can write the last state and the SET and RESET lines can be held low while the PROM is written. This is advantageous because it lowers the number of I/O pins required for a given write time. The shadow registers can be quickly written (sequentially) and then, when the light is turned on, one can transfer

their contents, in parallel, to the floating gates. A cell which does this is illustrated in Fig. 3. SE enables the shadow latch.

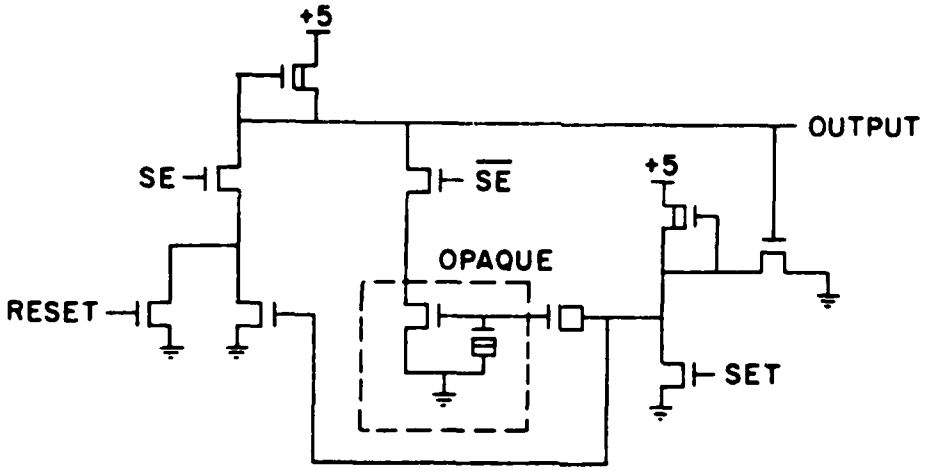


Fig. 3 Schematic of PROM with shadow memory.

The UV Write-enabled PROM is new field programmable device that allows one to add a few hundred bits of non-volatile storage to virtually any silicon gate MOS chip, with no additional masking or processing steps. There are many systems applications for this capability. Typical applications might include the storage of cryptographic codes, special addresses, calibration data, or repair locations for fault-tolerant systems.

The long term reliability of these devices is presently under investigation. We have demonstrated months of storage time at room temperature. Since the storage mechanism is precisely the same as that used in many commercial UV-PROMs, we anticipate few problems. The oxides which blanket the floating gate are either thermally grown or very thick. Previous work on floating gate MOS structures is highly encouraging in this regard [1, 2].

Figure 4 shows a photomicrograph of the experimental cell. The layout was not optimized. We now understand that the grounded metal line shading v_{bit} should have been attached to the OUTPUT instead of ground because it provides a source of photo-excited electrons near the floating gate, thereby limiting the gate's maximum potential. Since the floating gate is not charged to V_{DD}, a low trigger voltage on the SET NOR gate is required. This can be accomplished by making the pullup long.

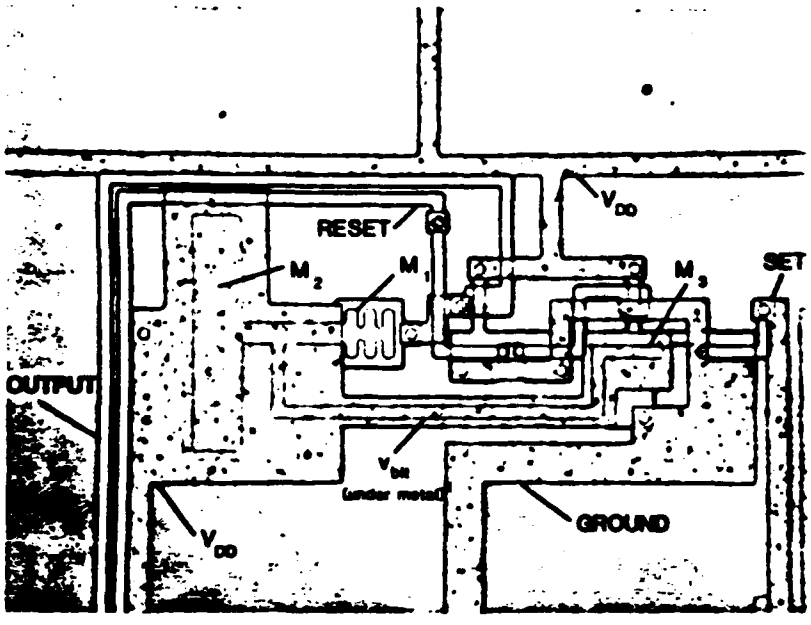


Fig. 4 Photomicrograph of experimental cell.

It is my pleasure to acknowledge the assistance of Mr. Paul Bassett and Mr. Jain Li. An enlightening conversation with Professor Steve Senturia is also gratefully acknowledged.

References

[1] D. Frohman-Bentchkowksy, "A fully-decoded 2048-bit electrically programmable FAMOS read-only memory," *IEEE J. Solid State Circuits*, vol. SC-6, pp. 301-306, 1971.

[2] D. C. Shaver, "Electron Beam Testing and Restructuring of Integrated Circuits," *Second Caltech Conf. on Very Large Scale Integration*, pp. 111-126, Pasadena, CA, 1981.

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-238 March 1985

Bounding Techniques and Applications for VLSI Circuit Simulation*

Charles A. Zukowski, John L. Wyatt,Jr., and Lance A. Glasser**

ABSTRACT

Simulation of large digital MOS circuits has recently been made faster both by using simplified device models and by using algorithms that are tailored for such circuits. A bounding approach is presented that builds upon these techniques to trade speed for measured uncertainty, allowing uncertainty to be managed efficiently. The bounding techniques can also be used to incorporate uncertainties in the circuit model arising from variations in fabrication process and operating conditions, and uncertainties in input waveforms. Monotonic properties of the MOS circuit model allow rigorous bounds to be generated through the analysis of simplified circuit models, and waveform relaxation techniques can be extended to efficiently include bounds for large digital MOS circuits. When applied at a high level, bounding algorithms can provide a useful enhancement for VLSI circuit simulation programs.

---

*To be presented at the 1985 IEEE International Symposium of Circuits and Systems. This work was supported by an IBM Graduate Fellowship, by the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622, and by the National Science Foundation under grant number ECS83-10941.

**Department of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139, Zukowski: Room 36-585, (617) 253-8169, Wyatt: Room 36-865, (617) 253-8169, Glasser: Room 36-587, (617) 253-4677.

Bounding Techniques and Applications for VLSI Circuit Simulation

Charles A. Zukowski, John L. Wyatt, Jr., and Lance A. Glasser

Abstract

Simulation of large digital MOS circuits has recently been made faster both by using simplified device models and by using algorithms that are tailored for such circuits. A bounding approach is presented that builds upon these techniques to trade speed for measured uncertainty, allowing uncertainty to be managed efficiently. The bounding techniques can also be used to incorporate uncertainties in the circuit model arising from variations in fabrication process and operating conditions, and uncertainties in input waveforms. Monotonic properties of the MOS circuit model allow rigorous bounds to be generated through the analysis of simplified circuit models, and waveform relaxation techniques can be extended to efficiently include bounds for large digital MOS circuits. When applied at a high level, bounding algorithms can provide a useful enhancement for VLSI circuit simulation programs.

I. Introduction

The size and complexity of VLSI circuits has created a need for more powerful electrical circuit simulation tools [1]. In recent years research towards this goal has fallen into two major categories. First, since today's VLSI circuits are primarily digital MOS, simulation algorithms have been investigated that exploit the special properties of these circuits to gain efficiency. The latency of digital circuits has been exploited to eliminate unnecessary calculations corresponding to inactive portions of a circuit. Analysis has also been partitioned to achieve a linear scaling of computation time with circuit size. Second, since high accuracy is not always essential, algorithms have been explored that trade accuracy for speed. Macromodeling of basic cells, and linear and piecewise linear device models have been used to reduce complexity.

Although algorithms tailored to digital MOS circuits have been successful in reducing computation time, they eventually reach a limit. Trading accuracy for more speed, when feasible, also provides an attractive tool that is often used in conjunction with digital MOS algorithms. Accuracy can be sacrificed in many instances without any detrimental effect, as in the analysis of non-critical logic paths. Unfortunately, uncertainty cannot be managed effectively in standard approximating simulators because it is not measured. Many circuit designers continue to seek the reassurance of an "exact" simulator at great cost. Rough timing information can be very

useful in early design phases, but is not sufficient for high performance design. If a combinational logic block were being designed that had to be faster than 10ns, a simulator that estimated a 9ns delay for its critical path would not be helpful unless its standard deviation were only a few percent. If the number of paths with a delay close to 9ns were large, even more statistical accuracy would be required to achieve high confidence that the specification was met. If bounding algorithms were used, only enough computation would be required to achieve a ten percent level of accuracy for the critical paths, regardless of the number considered.

This paper presents an overview of our recent work on bounding techniques for VLSI circuit simulation. Space considerations prevent discussion of the technical details here, but a complete presentation is given in [2]. The main motivation to consider bounding techniques in VLSI circuit simulation is to further improve efficiency through uncertainty management. By trading measured accuracy for speed, only a required level of accuracy need be achieved. Fortunately, present algorithms for both efficient and approximate simulation of large digital MOS circuits can be extended to consider efficient bounds. Bounding can be viewed as a framework in which these tools can be applied rationally to a particular simulation problem.

In addition to measuring uncertainties that arise from simplified analysis, bounds can be used to incorporate uncertainties in the circuit model arising from variations in fabrication process and operating conditions, as well as uncertainties in input waveforms. In this manner "circuit simulation" can be extended into the realm of "circuit analysis," in which an entire class of circuits, corresponding to ranges of device models, etc., is considered simultaneously. For example, bounds can incorporate the groups of excitations considered in input-independent analysis.

When considering general problems, it is often difficult to generate tight bounds. Digital MOS circuits, though, provide a special problem in which the difficulties exhibited in more general problems can be avoided to a large extent. The second section discusses the feasibility of using bounding in VLSI circuit simulation in general terms.

In addition to being generally feasible, bounding algorithms for digital MOS circuits can use modifications of existing efficient approximation algorithms. Due to special monotonic properties of the MOS circuit model, simple linear or piecewise linear circuit models can be derived whose behavior bounds that of the original. By using algorithms developed to analyze these simple circuit models, rigorous upper and lower bounds on signal waveforms can be generated, producing approximations with bounded uncertainty. The third section discusses the use of

simplified models to generate bounds.

When considering a digital MOS circuit model with feedback, including local feedback produced by Miller capacitance, even simplified bounding models become quite complex. To generate tight bounds, simplified models must also contain the feedback paths present in the original circuit. The techniques developed for efficient exact analysis of such circuits can also be extended to include bounds. More specifically, the Waveform Relaxation algorithm [3] can be extended to include waveform intervals, and the behavior of small tightly coupled subcircuits can be bounded separately. The partitioning essential for efficient VLSI simulation can still be achieved. The performance of the algorithm can be roughly maintained in the bounding context, and bounds can be generated for even sophisticated circuit models with a computation time that scales roughly linearly with circuit size. The extension of Waveform Relaxation to include bounds is discussed in the fourth section.

Experiments have shown that useful bounds can be constructed for many of the special subcircuits found in digital MOS logic. These results are discussed briefly in the fifth section. As more sophisticated bounding algorithms are developed, they should provide a useful enhancement to VLSI simulation programs. Bounding algorithms have the nice property that they never give an incorrect answer. In the worst case, a bounding algorithm might not provide sufficient accuracy in a given subcircuit at an acceptable cost, forcing the use of conventional exact algorithms. As more efficient algorithms are developed for a wider range of subcircuits, they can be incorporated incrementally to improve average performance.

II. Digital MOS Circuits

The behavior of digital MOS circuits can be successfully bounded at the level of logic signal waveforms, but in other problems bounding has proven more difficult. A simple example can illustrate why some calculations are difficult to bound efficiently. Consider first the calculation $Y = X - X$. If the variable X is known to lie in the interval [0,1], a straightforward bound on the subtraction operator will conclude that Y must lie in the interval $[-1,1]$. By ignoring any "correlations" between the two operands of the subtraction, calculations are made feasible but information is lost. Y must be zero even if X is completely unknown. In this case ignored correlations amplify uncertainty.

A calculation that does not exhibit a correlation problem is $Y = X + X$. Knowledge that the two operands must be identical does not change the conclusion that Y must lie in the interval [0,2] if X lies in [0,1]. Uncertainty in the delay of restored signal waveforms is more nearly analogous to the

adder example. Rough bounds on delays have been used successfully by TTL circuit designers for years. The main correlation that is ignored in digital logic circuits is that among logic signals. Roughly speaking, extreme circuit behaviors arise when signals and devices are either all slow or all fast. Unrealistic situations in which some are slow and others fast produce behaviors that fall somewhere in between. As a result, ignoring this correlation does not generally amplify uncertainty.

Basic uncertainty in a simulation arises from the use of incompletely specified models, the use of simplification to speed computation, and the use of bound relaxation sequences not taken completely to convergence. As long as ignored correlations do not amplify this basic uncertainty to a large extent, it can be kept quite small even when substantial simplifications are made.

We have found that there are additional correlations that must be ignored to produce simple and rigorous bounding algorithms for complex digital MOS circuit models, but these involve second order effects for restoring logic circuits. Examples are correlations between signals and their derivatives, correlations between values of derivatives over time, and correlations between the effects of coupling elements such as Miller capacitors on the two subcircuits they connect. The ignored correlations only have a large effect for certain types of MOS subcircuits. For example, bootstrap drivers depend on strong correlations among variables to operate correctly. Also, circuits such as ring oscillators with strong negative feedback are sensitive to correlations among signals, even when considered only for a small number of cycles. Dynamic nodes with capacitive coupling need bounds on total charge sharing to maintain voltages in the face of correlations over time. For standard restoring logic circuits, though, immunity to uncertainty amplification makes bounding techniques very attractive. Only for the small portion of a circuit containing difficult circuits is virtually exact analysis required, and these can potentially be found with a bounding algorithm by observing where uncertainty is amplified.

III. Simplified Bounding Models

The monotonic properties of the MOS circuit model can be used to bound its response through the simulation of a simplified model. We consider monotonicity of circuit behavior with respect to both input waveforms and element constitutive relations. A monotonic function maps bounds on its operands into rigorous bounds on its output, and nonlinear input waveforms and element constitutive relations can be tightly bounded by piecewise linear ones. As a result, the behavior of any circuit with the desired monotonicity can be bounded using piecewise linear analysis. Computation costs are similar when using approximate piecewise linear analysis, but uncertainty can be measured when bounds are used.

It is assumed here that a VLSI circuit model consists of discrete resistors, capacitors, n-channel transistors, and p-channel transistors with continuous, monotonic constitutive relations. The circuit is excited with external, grounded voltage sources. The transistor models considered are purely resistive, so all internal transistor capacitance is modeled with capacitors at the transistor terminals. Most rough or sophisticated MOS circuit models currently used fall into this class of networks.

To use waveform relaxation algorithms along with nodal analysis, it is sufficient to assume that all nodes with capacitance are connected to each other through some path of finite, positive incremental capacitance. In an integrated circuit model where every node is connected to a grounded capacitor, this condition is satisfied. For networks that do not satisfy this constraint, arbitrarily small capacitors must be added. The nodes in the capacitor network that are also connected to non-capacitive elements are called "independent" nodes, and their voltages are used to represent circuit behavior.

In this section, only the behavior of small subcircuits is considered, as relaxation can be used to combine these to analyze the behavior of the entire network. Consider first a Waveform Relaxation partitioning [3] that corresponds to isolating each of the independent nodes, and using the behaviors of the other nodes as inputs. In this case, we have been able to prove [2] that the behavior of each subcircuit is monotonic in its input waveforms and in the constitutive relations of its elements if the following conditions hold:

- Only independent nodes are connected to two or more elements of different type, e.g., resistors and n-channel transistors.
- All elements form series-parallel "one-port" groups that directly connect two independent nodes. (One-ports are defined for transistors as if they were resistors from drain to source.)
- All transistor gate and substrate terminals must be connected to independent nodes, except for the special case of depletion load pullups.

The most sophisticated MOS circuit models, containing some finite grounded capacitance at every node, automatically satisfy all of these constraints. The constraints are unrestrictive enough that most intermediate complexity models that ignore or group various capacitances are also included. Any circuit that violates one of the constraints can be transformed into a slightly more complex legal one with the addition of arbitrarily small grounded capacitors and arbitrarily large grounded resistors. As a result, the solution for each independent node behavior in a MOS integrated circuit can be bounded by the solution of a much simpler circuit, e.g., one containing

piecewise linear input waveforms and devices.

Partitioning of MOS circuits usually is more efficient when taken only to the level of strongly connected clusters. A cluster is defined roughly as a group of nodes connected by a d.c. current path that does not pass through the power supply. To extend the monotonic relationships to larger partitions, a circuit transformation must be used. Any element that connects two nodes within the partition must generally be duplicated to separate its effect on each node. A grounded dependent source is connected to each instance of the element to represent the effect of the other node. After such a transformation, the cluster solution becomes a monotonic function of both its input waveforms and element constitutive relations, if duplicated elements are treated as distinct. For many cluster circuits that contain transistors, the correlations that are ignored as a result of this transformation can represent a significant source of information loss.

Many special cluster circuits that appear often in simplified circuit models have more powerful monotonic properties that have been investigated. Many elements do not need to be duplicated to guarantee monotonicity, removing an ignored correlation and further simplifying the bounding circuit at the same time. One important case that we have considered is an RC tree driven by a restoring logic gate, where the resistors are fixed, there is no internodal coupling capacitance, the pullup and pulldown networks have no internal capacitance, and the output is monotonic in time [4]. The special properties of nonlinear RC lines, meshes, and trees with various restrictions have also been considered in detail [5], [6], [7]. If the response of a cluster can be bounded by that of a linear RC circuit, results exist that allow the response of the linear circuit to be bounded with simple closed-form expressions [8], [9].

IV. Bound Relaxation

The key to efficient analysis of digital MOS circuits is partitioned analysis. For intermediate complexity circuit models with no feedback through MOS transistors, event driven simulation has been used along with a straightforward partitioning into clusters. Bounding the response of such circuit models can take advantage of the same techniques in a straightforward manner, propagating bounds from inputs of clusters to outputs.

The Waveform Relaxation algorithm is an example of a partitioning method for general circuit models that performs well in exact simulation. Each tightly coupled subcircuit, corresponding roughly to a logic block, is analyzed separately over large time intervals to iteratively improve estimates for the solution. Latency is exploited by using variable time steps across subcircuits, convergence is fast for most digital MOS circuits, and computation cost scales roughly linearly

with circuit size.

Waveform Relaxation can be extended to handle intervals of waveforms with performance similar to that of the exact algorithm [10]. As a result, the efficiency and linear scaling arising from partitioning can be used when bounding complex circuit models. For bounding applications, the advantage of restricting analysis to small subcircuits appears to be large, as the complexity of generating direct bounds appears to grow very quickly with circuit size.

The exact Waveform Relaxation algorithm is based on a "relaxation function" that maps one estimate of circuit behavior to a better one based on the exact response of each subcircuit. We consider a "bounding function," a map from intervals of circuit behavior to new intervals based on bounds on the responses of each subcircuit. The results in the previous section show how bounding functions can be calculated efficiently. The bounding function is defined so that its output interval contains the image of its input interval under a relaxation function. We have been able to show that such a bounding function has two important properties:

- A bounding function maps a rigorous bound on the solution into another rigorous bound.
- A tentative bound on the solution which is mapped to a tighter one by a bounding function is guaranteed to be a rigorous bound.

A tight bounding function roughly maintains the contraction property of the exact algorithm upon which it is based. If a conservative guess for the solution can be generated, a bounding function can be used both to check if it is a valid bound and to tighten it through repeated relaxation. Relaxation using a bounding function will tend to converge quickly to an interval that is self consistent, i.e., the bounding function maps the interval to itself. If the sequence of intervals is telescoping from a very conservative initial guess, any element in the sequence is a valid bound. If the sequence is taken to convergence, the solution corresponds to the direct solution obtained by ignoring correlations between subcircuits.

V. Experimental Results

Some experimental programs have been written to assess the feasibility of a bounding approach. Algorithms appropriate for intermediate complexity models of gate arrays have measured delay uncertainty in the range of $\pm$30% with simplifications corresponding to using RC bounding circuits for each gate [4]. Accuracy can be increased by using piecewise constant transistor bounds with any number of segments. The cost of measuring the uncertainty in approximating gate array simulation algorithms is not large, and a bounding algorithm can find

critical paths and produce tight bounds on the outputs without wasting computation where it is not necessary.

A program to relax bounds on general linear RC mesh circuits has also been developed [11]. Since the relaxation can be calculated exactly in the linear case and feedback is positive between subcircuits, precise knowledge of the element values and initial state leads to a sequence of closed-form bounds that becomes arbitrarily tight. The algorithm can be used to tighten direct closed-form bounds on RC trees for circuits in which they are loose, or handle "leaky" RC meshes with arbitrary initial state not treated by direct bounding algorithms. Experiments using the program have indicated that even within small clusters, relaxation techniques can have acceptable performance, producing delay bounds that are tighter than conventional ones [8] after only a few iterations.

Lastly, simple bounding algorithms for general MOS circuit models have been tested on combinational logic circuits containing Miller capacitors. The bounding algorithm for subcircuit responses uses piecewise constant transistor models and produces piecewise linear voltage waveforms. Using simplified models that generate delay uncertainty on the order of $\pm 5\%$ for circuits without Miller capacitance, the addition of Miller capacitance only increases delay uncertainty by roughly a factor of two or three. In addition, at the $\pm 5\%$ level of accuracy, bound relaxation exhibits convergence behavior similar to that of exact relaxation.

VI. Conclusion

A theoretical framework has been established that should lead to efficient bounds on the behavior of a large portion of typical digital MOS integrated circuits. A bounding approach to simulation has many advantages, and initial experiments regarding feasibility have been very encouraging. More experience with a simulation program based on bounding techniques is needed to guide further research in the field. Towards that end, work is required on chosing detailed strategies for bounding simulators and high level algorithms for managing uncertainty once it can be measured. More powerful results are also desirable for general cluster circuits, important special subcircuits, and simplified models.

Acknowledgements

We would like to acknowledge useful discussions with Paul Penfield, Jr. This work was supported by an IBM Graduate Fellowship, the Defense Advanced Research Projects Agency under contract number N00014-80-C-0622, and the National Science Foundation under grant number ECS83-10941.

References

[1] Albert E. Ruehli and Gary S. Ditlow, "Circuit Analysis, Logic Simulation, and Design Verification for VLSI," *Proceedings of the IEEE*, Vol. 71, No. 1, January 1983, pp. 34-48.

[2] Charles A. Zukowski, *Bounding Enhancements for VLSI Circuit Simulation*, PhD dissertation, MIT, 1985.

[3] Ekachai Lelarasmee, Albert Ruehli, and Alberto Sangiovanni-Vincentelli, "The Waveform Relaxation Method for Time-Domain Analysis of Large Scale Integrated Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 3, 1982, pp. 131-145.

[4] Charles A. Zukowski and Lance A. Glasser, "Bounding the Response of One-Way MOS Logic Gate Models," VLSI Memo 84-176, MIT, May 1984.

[5] John L. Wyatt, Jr., "Monotone Sensitivity of Nonlinear Nonuniform RC Transmission Lines, with Application to Timing Analysis of Digital MOS Integrated Circuits," *IEEE Transactions on Circuits and Systems*, Vol. CAS-32, No. 1, January 1985, pp. 28-33.

[6] Mark Horowitz, "Timing Models for MOS Pass Networks," *Proceedings of International Symposium on Circuits and Systems*, May 1983, pp. 198-201.

[7] Qingjian Yu and Omar Wing, "Waveform Bounds of Nonlinear RC Trees," *Proceedings of IEEE International Symposium on Circuits and Systems*, IEEE, May 1984, pp. 356-359.

[8] Jorge Rubinstein, Paul Penfield, and Mark Horowitz, "Signal Delays in RC Tree Networks," *IEEE Transactions on Computer Aided Design*, Vol. CAD-2, No. 3, July 1983, pp. 202-211.

[9] John L. Wyatt, Qingjian Yu, Charles A. Zukowski, Han-Ngee Tan, and Peter O'Brien, "Improved Bounds on Signal Delay in MOS Interconnect," VLSI Memo 84-221, MIT, January 1985.

[10] Charles A. Zukowski, "Circuit Simulation with Iterating Bounds," VLSI Memo 84-159, MIT, January 1984.

[11] Charles A. Zukowski, "Relaxing Bounds on Linear RC Mesh Circuits," VLSI Memo 84-166, MIT, April 1984.

IMPROVED BOUNDS ON SIGNAL DELAY IN LINEAR RC MODELS FOR MOS INTERCONNECT

Qingjian Yu[2], John L. Wyatt, Jr.[1], Charles Zukowski[1],
Han-Ngee Tan[3], Peter O'Brien[1]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[2]Electrical Engineering Teaching Office, East China Institute of Technology, Nanjing, Jiangsu, China

[3]School of Electrical and Electronic Engineering, Nanyang Technological Institute, Upper Jurong Road, Singapore 2263.

Abstract

Computationally simple bounds for signal propagation delay in linear RC tree models for MOS interconnect were derived in [1] and have proved useful in timing analysis of digital MOS IC's [2-4]. We show that these bounds can be derived quite simply as the payoff functions for a certain linear optimal control problem and that they apply not only to RC trees but to more general RC meshes as well. Finally, two methods are given for tightening the original bounds given in [1].

I. Introduction

In digital integrated circuits, signal propagation delay through conducting paths grows in relative importance as feature sizes shrink. Bounds on the delay, applicable to those paths that can be modelled as "RC trees," were derived in [1]. But, as discussed in [5-7], certain circuits used in MOS logic cannot be modelled as RC trees because they contain one or more loops of resistors, as shown in Fig. 1. Several examples of such circuits, called "RC meshes," arising in MOS logic networks are given in [5,6].

This paper is concerned with bounds on the zero-state step response and hence on signal propagation delay in linear, lumped RC tree and mesh networks driven by an ideal voltage source.

II. Network Differential Equations for RC Meshes

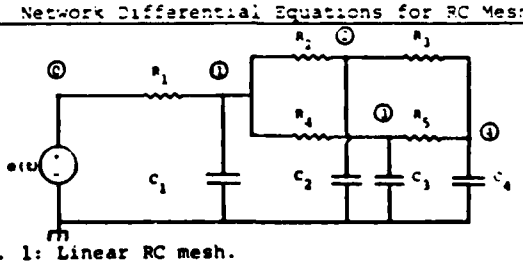


Fig. 1: Linear RC mesh.

Isolate the resistor subnetwork $\bar{R}$ containing all the resistors and assign reference directions to the capacitor currents as shown in Fig. 2. The node voltages with respect to datum are given in terms of the positive-definite, symmetric matrix $\bar{R}$ as shown below.

$$\begin{pmatrix} v_1 - e \\ \vdots \\ v_N - e \end{pmatrix} = \begin{bmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & & \vdots \\ r_{N1} & \cdots & r_{NN} \end{bmatrix} \begin{pmatrix} i_1 \\ \vdots \\ i_N \end{pmatrix} \qquad (1)$$

Consider the step response. Substituting $e=1$ and $i_k = -C_k \dot{v}_k$ into (1), yields the network differential equations

$$1 - v_i(t) = \sum_{j=1}^{N} r_{ij} C_j \dot{v}_j(t), \quad i=1,\ldots,N, \; t > 0, \qquad (2)$$

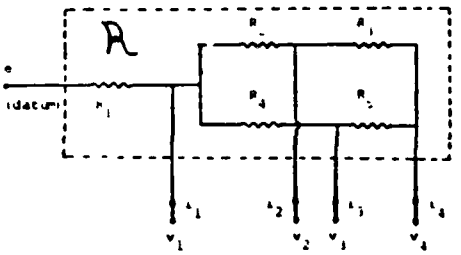


Fig. 2: The resistor subnetwork $\bar{R}$.

which are identical in form to eq. (9) of [1]: the only difference is that in [1] certain resistances R_{ij}, defined specifically in terms of the topology of a tree, appear in place of the r_{ij}'s above.

III. Optimal Control Method for Determining Bounds on the Step Response

The derivation outlined below is an alternative to that in [1] and yields essentially the same results, but it applies to meshes as well and also affords a natural way to incorporate additional information and thereby obtain tighter bounds, as shown in Section IV.

Fact 1

For any three nodes, i, j, k of an RC mesh,

$$r_{ii}\, r_{kj} \geq r_{ki}\, r_{ij} . \qquad (3)$$

The proof of (3), given in [6], generalizes the argument for the special case of a tree in [1].

CH2114-7/85/0000-0903$01.00 © 1985 IEEE

Fact 2

The zero-state step response of an RC mesh is completely monotone, i.e.,

$$\dot{v}_j(t) \geq 0, \quad j=1,\ldots,N, \quad \forall t \geq 0 . \tag{4}$$

The proof is in [8].

Fact 3

For any two nodes i and k of an RC mesh and any instant t during the step response,

$$r_{ii}(1-v_k(t)) \geq r_{ki}(1-v_i(t)) \tag{5}$$

$$r_{ki}(1-v_k(t)) \leq r_{kk}(1-v_i(t)) . \tag{6}$$

Given Facts 1 and 2, the derivation of (5) and (6) is identical to that in Appendix D of [1].

At this point the strategy becomes one of reduced order modelling with time-domain error bounds. Choosing a distinguished node i as the output node of interest, we describe the system in terms of only two state variables, the distance to equilibrium $(1-v_i(t))$ and its integral

$$f_i(t) \triangleq \int_t^\infty (1-v_i(t'))dt' = \sum_k r_{ik}C_k(1-v_k(t)) , \tag{7}$$

where the last equality follows upon substituting (2) into the integral and evaluating. Using (5) and (6) in (7) yields the following inequality between these two state variables:

$$\underbrace{\sum_k r_{ik}^2 C_k / r_{ii}}_{\triangleq T_{P_i}} (1-v_i(t)) \leq f_i(t) \leq \underbrace{\sum_k r_{kk}C_k}_{\triangleq T_P} (1-v_i(t)), \quad \forall t \geq 0 . \tag{8}$$

From (7) one initial condition is

$$f_i(0) = \sum_k r_{ik}C_k \triangleq T_{D_i} . \tag{9}$$

It was shown in [1] that step response bounds can be obtained by appropriate manipulations of (4, 5) and (7-9) above, but the methodology is somewhat obscure. A clearer view emerges from recasting the calculations into the form of a linear minimum- (and maximum-) time optimal control problem with state constraints, in which an input u(t) is introduced to represent the unknown waveform $\dot{v}_i(t)$:

Minimize (or maximize) T

for the dynamical system

$$\dot{f}_i(t) = -(1-v_i(t)) \tag{10}$$

$$\frac{d}{dt}(1-v_i(t)) = u(t) , \tag{11}$$

with initial conditions

$$f_i(0) = T_{D_i}, \quad (1-v_i(0)) = 1 , \tag{12}$$

state constraints

$$T_{R_i}(1-v_i(t)) \leq f_i(t) \leq T_P(1-v_i(t)), \quad \forall t \geq 0 , \tag{13}$$

input constraint: $u(t) \leq 0, \quad \forall t \geq 0 ,$ (14)

and terminal condition

$$(1-v_i(T)) = (1-v_i^*), \quad 0 \leq v_i^* \leq 1 . \tag{15}$$

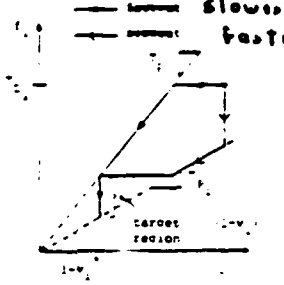


Fig. 3: Fastest and slowest trajectories.

The optimal trajectories can be determined by inspection, since the time duration of any path in the $(1-v_i)$ - f_i plane can be found by rearranging and integrating (10) to yield

$$T = \int_{f_{final}}^{f_{init}} \frac{1}{1-v_i} \, d f_i . \tag{16}$$

Thus the fastest trajectory from the initial point to the target interval is the one for which both the region of integration (f_{final}, f_{init}) and the integrand $(1-v_i)^{-1}$ are minimized, and the slowest trajectory is found similarly. See Fig. 3. The minimum and maximum times depend on the "target" voltage v_i^* and are denoted $T_{min}(v_i^*)$ and $T_{max}(v_i^*)$. The inverse functions, denoted respectively $\bar{v}_i(t)$ and $\underline{v}_i(t)$, are the upper and lower bounds, respectively, for the step response of the mesh. The algebraic form of $\underline{v}_i(t)$ and $\bar{v}_i(t)$ obtained in this way can be easily read off from Fig. 3 and agrees with the results in [1]: the exact expressions are omitted for the sake of brevity. They approach a well-defined limit in the case of a distributed network, e.g., the simple example in Fig. 4, for which they are plotted in Fig. 5.

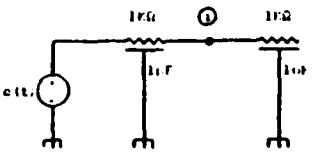


Fig. 4: The bounds approach a well-defined limit for a distributed network such as this one, for which $T_{P_i} = 1.33$ ns., $T_{D_i} = 1.5$ ns., $T_P = 2.0$ ns, and $T_{R_i} = 0.33$ ns.

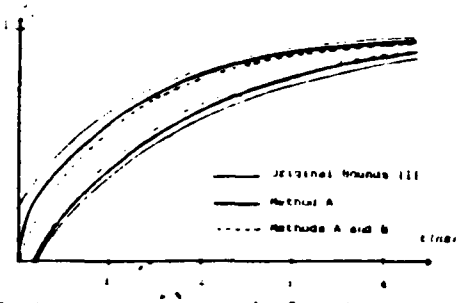


Fig. 5: Step response bounds for the network in Fig. 4, with output taken at node 1.

The bounds represent an effort to approximate the dynamics of a higher order network by one with a single time constant T_{D_i}: they are exact only in that case. Whenever $(T_{D_i} - T_{R_i}) \ll T_{D_i}$, the wedge-shaped region in Fig. 3 is quite narrow and the bounds will be quite tight. Chapter 3 of [7] gives examples of networks for which the bounds are good and others where they are poor.

IV. Method "A" for Bounds Improvement: Limits on the Maximum Slew Rate of Node Voltages

The optimal trajectories shown in Fig. 3 include horizontal segments along which v_i changes while f_i remains constant, corresponding to instantaneous jumps in v_i that cannot occur in practice. We can tighten the bounds by adding constraints eliminating such trajectories. The simplest form for such a constraint is a "minimum slope bound" in the $(1-v_i) - f_i$ plane of the form

$$\frac{d f_i}{d(1-v_i)} \geq \tau_i > 0 \ . \qquad (17)$$

This rules out both trajectories in Fig. 3 as feasible solutions. The new optimal trajectories are as shown in Fig. 6, and the corresponding algebraic form for $v_i(t)$ and $\dot{v}_i(t)$ is given in [9].

The inequality (17) corresponds to a "slew-rate bound," i.e., a bound on the derivative, for $\dot{v}_i$, since

$$\frac{\dot{v}_i}{(1-v_i)} = -\dot{(1-v_i)}/(1-v_i) = \dot{(1-v_i)}/\dot{f}_i = \frac{d(1-v_i)}{d\, f_i} \leq \frac{1}{\tau_i} . \qquad (18)$$

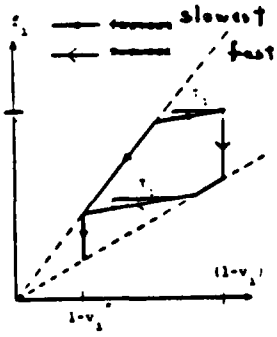


Fig. 6: Altered optimal trajectories.

For any mesh we know that $\tau_i \geq r_{ii}C_i$, since

$1-v_i = \sum_{j=1}^{N} r_{ij}C_j\dot{v}_j \geq r_{ii}C_i\dot{v}_i$, from (2) and (4).

Using $\tau_i = r_{ii}C_i$ can significantly tighten the bounds whenever the mesh contains only a small number of lumped capacitors, as is commonly the case in reasonably accurate circuit models for distributed interconnect [10]. But as progressively more R's and C's are used to model a given section of interconnect, $C_i \to 0$ and (17) becomes useless with $\tau_i = r_{ii}C_i$.

Fortunately, values of τ_i greater than $r_{ii}C_i$ can be found for many RC trees. Space constraints limit us to mentioning only one of the results in this direction obtained in [11]. Consider an RC line with the nodes numbered in increasing order as one moves away from the source. It was first noted in [7] that for such a network

$$\frac{\dot{v}_j(t)}{1-v_j(t)} \geq \frac{\dot{v}_i(t)}{1-v_i(t)} \ , \ \forall j \leq i, \ \forall t \geq 0 \ , \qquad (19)$$

a rigorous proof was given and the result further extended in [11]. Using (19) and (5) in (2), one can show that if i is any node of an RC line, or any node of an RC tree between the source and the first branch point, then

$$\tau_i \geq \sum_{j=1}^{i} r_{ij}^2 C_j / r_{ii} \triangleq T_{R_i} \ . \qquad (20)$$

This improvement is illustrated in Fig. 5.

V. Method "B" for Bounds Improvement: Spatial Convexity of Node Voltages

At any instant during the step response of an RC line or tree, the node voltages are a convex function of distance from the source. For the network in Fig. 4, a characteristic voltage profile is plotted in Fig. 7, along with the bounds (5) and (6).

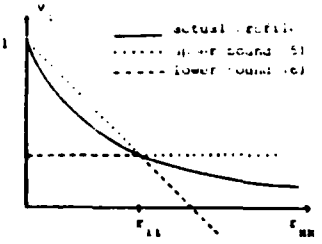


Fig. 7: Voltage profile for Fig. 4.

Considerable improvement over (6) is possible since a convex curve is bounded below by any tangent line, i.e.,

$$1-v_k \leq (1-v_i)\left[1 + \lambda\left(\frac{r_{kk}}{r_{ii}} - 1\right)\right] \qquad (21)$$

for some $\lambda \in [0,1]$. Substituting (21) into the right hand side of (7) and taking the maximum over λ yields

$$f_i \le (1-v_i) \max \{T_{D_i}, \sum_k \frac{r_{ik}r_{kk}}{r_{ii}} C_i\} \le T_p(1-v_i) , \quad (22)$$

thus reducing the effective value of T_p (from 2.00 ns. to 1.83 ns. for the network in Fig. 4) and further improving the voltage bounds as shown in Fig. 5. Current research includes extending this technique to trees.

Acknowledgement

It is a pleasure to acknowledge helpful conversations with Professors Lance Glasser and Paul Penfield of MIT, and Professor Mark Horowitz of Stanford University. This work was supported by the National Science Foundation under Grant No. ECS-8310941, the Air Force Office of Sponsored Research under contract No. F49620-84-C-0004, and the Defense Advanced Research Projects Agency under Contract No. N0014-80-C-0622.

References

[1] Rubinstein, J., P. Penfield, Jr., and M.A. Horowitz, "Signal Delay in RC Tree Networks," IEEE Trans. Computer-Aided Design, vol. CAD-2, no. 3, pp. 202-211, July 1983.

[2] Putatunda, R., "Autodelay: A Second-Generation Automatic Delay Calculation Program for LSI/VLSI Chips," Proc. IEEE Int. Conf. on Computer-Aided Design, Santa Clara, CA, November, 1984, pp. 188-190.

[3] Jouppi, N., "Timing Analysis for NMOS VLSI," ACM IEEE 20th Design Autom. Conf. Proc., June 1983, pp. 411-418.

[4] Tamura, E., K. Ogawa, and T. Nakano, "Path Delay Analysis for Hierarchical Building Block Layout System," ACM IEEE 20th Design Autom. Conf. Proc., June 1983, pp. 403-410.

[5] Wyatt, J.L., Jr., and Q. Yu, "Signal Delay in RC Meshes, Trees and Lines," Proc. IEEE Int. Conf. on Computer-Aided Design, Santa Clara, CA, November 1984, pp. 16-17.

[6] Wyatt, J.L., Jr., "Signal Delay in RC Mesh Networks," to appear in IEEE Trans. Circuits and Systems, 1985.

[7] Horowitz, M.A., "Timing Models for MOS Circuits," Technical Report No. SEL 83-003, Integrated Circuits Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305, December 1983.

[8] Wyatt, J.L., Jr., "Monotone Behavior of Non-linear RC Meshes," VLSI Memo No. 82-128, November 1982, p. 10. All VLSI Memos are available from the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139.

[9] Tan, H.N., and J.L. Wyatt, Jr., "Time-Optimal Trajectories Associated with Voltage Bounds in RC Tree Networks," VLSI Memo No. 84-205, October 1984.

[10] Sakurai, T., "Approximation of Wiring Delay in MOSFET LSI," IEEE J. Solid-State Ckts., Vol. SC-18, no. 4, August 1983, pp. 418-426.

[11] Yu, Q. and J.L. Wyatt, Jr., "Improved Bounds on Signal Delay in RC Trees Using Inequalities on the Derivatives of Node Voltages," VLSI Memo 84-197, August 1984.

Schema

An Architecture for Knowledge Based CAD

G. C. Clark*— R. E. Zippel‡

*Harris GSS, P. O. Box 37, Melbourne, FL 32902, USA
‡MIT Laboratory for Computer Science, Cambridge, MA 02139, USA

Abstract

Schema provides an integrated environment for all aspects of the synthesis and analysis of electronic designs from PC boards through circuit and mask design of VLSI devices. It simplifies the development of synthesis and analysis tools by using uniform data structures and by making available libraries of standard routines and advanced control structures appropriate for CAD tool development. Because all tools in the Schema environment utilize the same abstract data structures it is easy for tools to interchange data about a design or even peices of the design itself. Schema also permits much of the design to be done in a technology independent fashion by allowing the designer to delay implementation decisions until the last possible moment. The information associated with a particular component of a design is organized as a module. Modules contain schematics, icons, topologies, layouts, simulation results, and other descriptive information for this component. The descriptions contained in modules are implemented as procedures which utilize other modules in a hierarchical fashion. Schema is under joint development by MIT and Harris Corporation.

SCHEMA is an environment for developing knowledge based, computer aided design tools for electronic systems. The three major goals of its design are:

- Provide an integrated environment for all aspects of the synthesis and analysis of electronic designs from PC boards through circuit and mask design of VLSI devices.
- Simplify the creation of computer aided design tools by encouraging and supporting their construction from libraries of standard routines, by using uniform data structures and by providing libraries of advanced control structures appropriate for CAD development.
- Allow the designer to delay making decisions until necessary; for example, the technology (TTL, ECL, gate array or custom MOS) used in a logic design need not be specified until timing simulations or physical design is begun.

The key to achieving these goals is the development of a totally *integrated design environment* where design tools easily communicate and cooperate. This has been achieved by the innovative software architecture used in the development of SCHEMA.

SCHEMA achieves coherence not by specifying the interchange formats to be used between different CAD programs, but rather by specifying the data structures the programs should use. SCHEMA specifies a set of abstract data types for dealing with electronic designs, and a set policies to be used when dealing with the new data types. This approach provides a common layer on which different CAD tools may built, and it allows the CAD tools to invoke each other and easily cooperate by interchanging pieces of electronic designs.

These data types are implemented using an object oriented programming system called Flavors[4]. These structures represent circuit topologies and schematics, mask artwork, floorplans and simulation waveforms (both digital and analog). Circuit topologies represent the connectivity of a circuit; schematics are represenations of the graphic images of a circuit that are drawn on paper. Since these structures are instances of flavors, they also incorporate pieces of code that allow them to directly provide procedural functionality. That is, a transistor contains the information and *code* required to display itself on the screen, write itself out to a file or participate in a simulation. This raises the semantic level at which the CAD tools deal with objects, simplifying their development. It also allows implementation and operation decisions to be delayed and even changed without modifying the code that makes use of them.

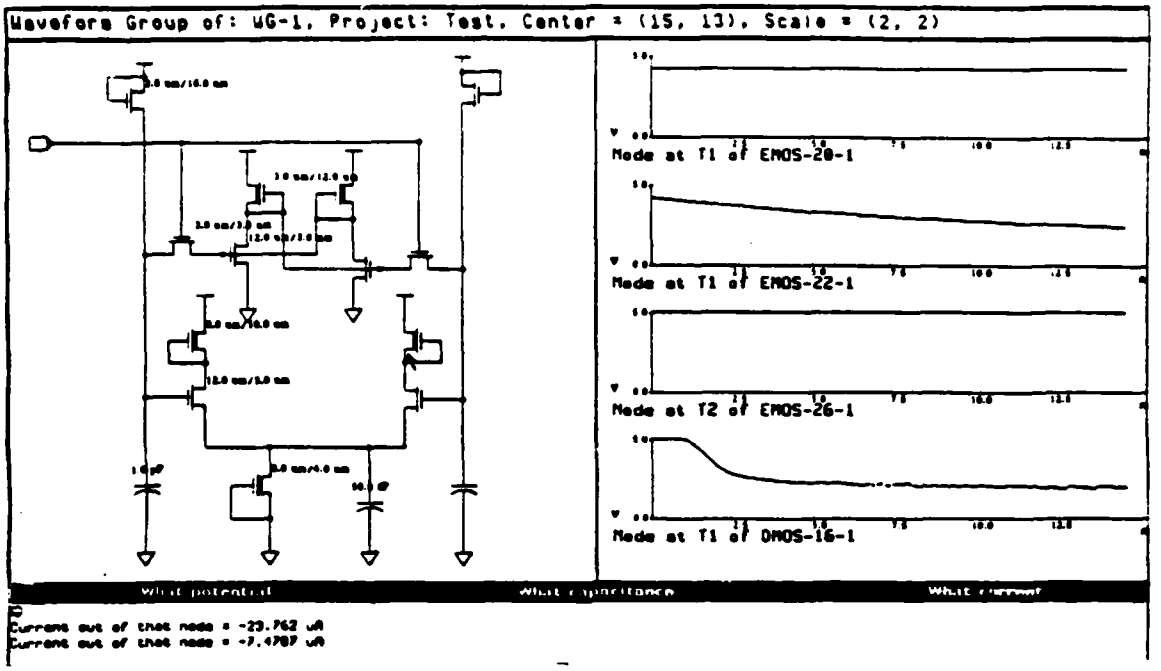
Modules

The basic component of a design in SCHEMA is a *module*. Each module consists of a *topology* and several *descriptions*, e.g. schematics, icons, layouts and simulation results. Examples of modules in a design include: an inverter, a half adder, an arithmetic logic unit, a data path, a cache, instruction fetch unit and a memory system. Each of these module includes not only the schematic (and its corresponding topology), but also the results of various tests that have been performed on the circuit (simulation results), documentation and design notes and physical specifications (VLSI layouts or PC board designs). The modules represent a complete view of a design component.

The designer rarely interacts directly with the topology of a module, but instead deals with the descriptions (schematics). The analysis tools (simulators, timing verifiers and other consistency checkers) work with the topology, and usually use the descriptions only for communicating with the designer. The only major exceptions are the physical design tools, VLSI layout system and wire wrap and PC board systems that, by necessity, must work with the physical descriptions.

The system ensures that the topology remains consistent with the descriptions provided by the designer through the use of timestamps and limited edit trails. It also warns the designer when two descriptions of a design become in-

Clock 1 2.psd



consistent. This division allows the electronic designer to use the most appropriate mechanism for describing the design without worrying about getting formats correct for the CAD tools, and the CAD tool designer deals only with design descriptions that are both appropriate and "pre-parsed."

The topology and its descriptions are implemented as procedures, though they are usually edited via one of the *description editors*: schematic, layout or waveform. This procedural structure, similar to the approach used in DPL[1], allows a great deal of flexibility parameterizing the different components and provides an excellent point at which to install intelligent synthesis modules. For instance, in an earlier version of SCHEMA this was used to implement an ALU module that chose different carry look-ahead schemes depending on the width of the data word[2].

These hierarchical descriptions also incorporate a *multiple viewpoint* or *Slices*[3] mechanism to allow simulation and analysis modules to annotate the topologies. The multiple viewpoints are used to control the visibility of certain information to the CAD tools. For instance, transient analysis programs like SPICE want to be aware of parasitic capacitances and resistances while a simple logic simulator might not. Rather than generating the two different topologies for the different simulators, the same topology is used for both but the parasitics are only visible when the *transient* viewpoint is made visible by SPICE. This way annotations to the topology made by the two simulators can be examined by their conterparts easily.

Project/Module Hierarchy

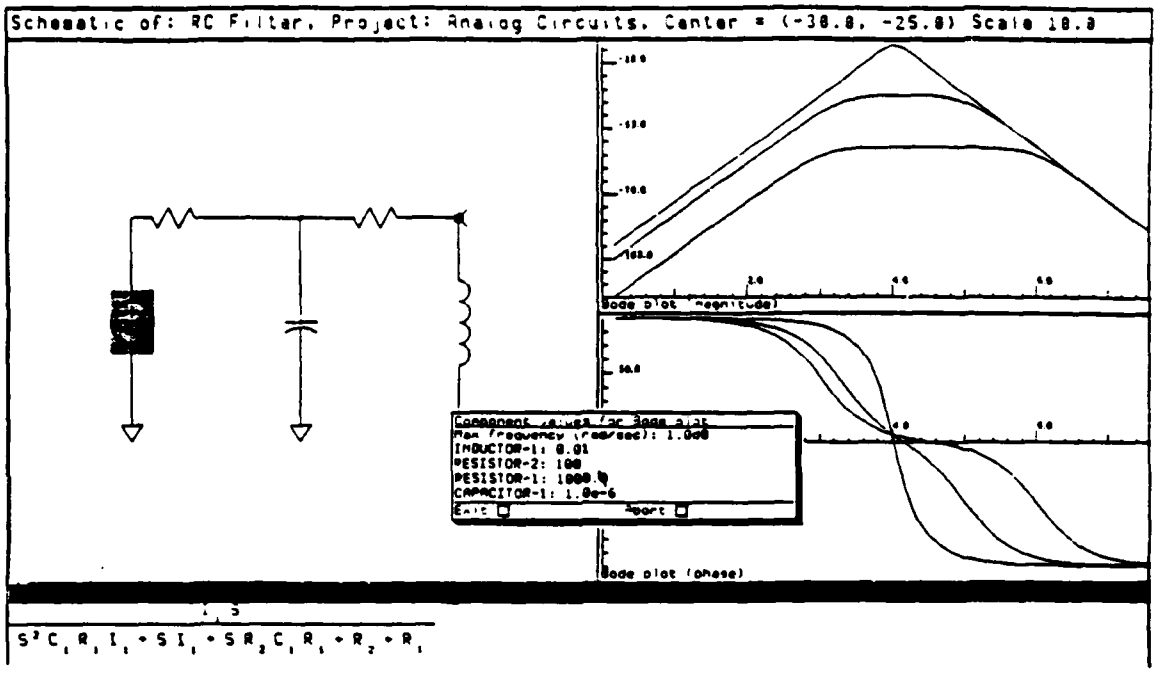
The modules and all other information relating to a design are collected into a *project*, which in turn can be a component of a larger project. For instance, there might be an *L machine* project that is used to hold all the design components of the L machine. Several different versions of the L machine might be designed, so there might be *TTL*, *CMOS* and *ECL* sub-projects of *L machine*. Within the *CMOS* project there might different projects to contain the design of the datapath, control logic, and memory management system. The modules of each of these projects would be combined by the main module contained in *CMOS* to produce the final chip.

Each designer maintains his or her own hierarchy of projects. The root of this hierarchy is called a *portfolio*. By having sub-projects point to the same *save file*, designers can share projects. This project/module hierarchy is a very useful way of organizing and managing the material related to a design.

Environments

By specifying an *environment* the designer makes precise what types of modules and tools should be available for the design. Each environment consists of a collection of primitive modules that may be used, command dispatch tables for the description editors, design rules, simulation models and so on. The environments themselves are organized as a directed acyclic graph. At any time, the designer can refine the environment being used. For instance, one could begin a design in the *Basic Logic* environment and later when it had been decided to use CMOS, switch to the *Generic CMOS* environment. Finally, when a foundry had been chosen, the designer would select an evironment for the specific process to be used. While the environment was *Basic Logic*, the designer would be able to draw logic schematics and simulation, but would be unable to get any timing information (other than in gate delay units) or do any circuit design. After switching to *Generic CMOS*, transistor level circuits and sticks diagrams could be developed. When the process specific environment has been chosen, detailed masks could be designed and accurate timing information would be available.

Circuit & 2 appl



Software Tools

Though most of the time the data structures needed by a CAD designer are already in place, SCHEMA also includes a large library of compatible flavors (abstract data types) for constructing new structures. Within this library are mechanisms for dealing with many different types of hierarchy, prototypes, "creation on demand," timestamping, and so on. When creating a new data structure, the designer merely picks the flavors that provide the functionality desired and includes his own customizations. This fine grained modularity has helped maintain a high level of uniformity within the system. The modularity techniques used are based on the Capsule ideas[5].

In addition there is also a growing library of useful CAD oriented procedures that may be drawn upon. Among them are sparse matrix routines, linear and non-linear equation solvers, a moderate size symbolic algebra package, topological traversal routines, two dimensional spatial management packages and so on. The existence of these packages has enabled CAD builders to build on each others work more than in previous systems.

The totality of these tools, mechanisms and policies remove much of the drudgery from CAD tool development and encourages tool developers to proceed in a cooperative, cummulative fashion. For the electronic designer it provides a uniform environment, with uniform access to a wide variety of different synthesis and analysis tools.

A simple transient simulator was built on this base by Chris Terman. An example of its use is shown on the preceding page. The results of the simulation are left as annotations on the topology that the user (or another program) can examine. The top left window shows the circuit being simulated, the top right one shows a few selected waveforms. In the bottom window, the currents into the depletion transitors are given.

The second figure illustrates the use of the linear systems analysis tools. The bottom window gives the exact transfer function of the *RLC* circuit shown in the top left window. At the right, several Bode plots are given, and a pop up menu is shown which gives the parameters of last plot.

Conclusions

We have given a brief summary of the internal architecture of SCHEMA and shown a few of its uses. Its novel architecture and extremely high degree of integration make SCHEMA easir to use both by CAD tool designers and designers than many other systems.

The authors would like to acknowledge the assistance of the other members of the SCHEMA design group, Anuja Kohli, Siu-Ling Ku, Mike McDonald, Margaret St. Pierre and Steve Seda. This work was supported by DARPA contract N00014-80-C-0622 and Harris corp.

References

[1] J. Batali and A. Hartheimer, "The Design Procedure Language Manual," MIT Artificial Inteligence Laboratory Report 598, 1980.

[2] M. Rose, *A Datapath Generator*. B. S. Thesis, Dept. EECS, Massachusetts Institute of Technology. June 1982.

[3] G. J. Sussman, "SLICES: At the Boundary between Analysis and Synthesis." *Proc: Artificial Intelligence and Pattern Recognition in Computer Aided Design*, IFIP WG 5.2, Grenoble, France, 1978.

[4] D. L. Weinreb and D. A. Moon, *Lisp Machine Manual*. MIT Artficial Intelligence Laboratory, Cambridge, MA, 1981.

[5] R. Zippel, "Capsules," *SIGPLAN Bulletin*, vol. 18, no. 6, pp. 166–169, 1983.

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE MASSACHUSETTS 02139

VLSI Memo No. 85-257 September 1985

Step Response Bounds for Systems Described by M-Matrices, with Application to Timing Analysis of Digital MOS Circuits*

J. L. Wyatt, Jr., C. A. Zukowski, and P. Penfield, Jr.**

ABSTRACT

Methods adapted from optimal control are used to calculate simple closed-form upper and lower bounds for the step response of linear systems governed by M-matrices. For high order systems these bounds can be calculated using far less computer time than an "exact" numerical solution requires. The technique has proven to be of real industrial significance in integrated circuit CAD, where it is used for rapid calculation of signal propagation delay through the myriad of interconnect paths in digital MOS IC's.

---

*To appear in Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, Florida, December 11-13, 1985. This research was supported by the National Science Foundation under Grant No. ECS83-10941 and by the Defense Advanced Research Projects Agency under Contract No. N00014-80-C-0622.

**Wyatt and Penfield: Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139; Wyatt: Room 36-864, (617) 253-6718; Penfield: Room 39-321, (617) 253-2506. Zukowski, current address: Department of Electrical Engineering, 1331 S.W. Mudd Bldg., Columbia University, New York, NY 10027.

STEP RESPONSE BOUNDS FOR SYSTEMS DESCRIBED BY M-MATRICES, WITH APPLICATION TO TIMING ANALYSIS OF DIGITAL MOS CIRCUITS

J.L. Wyatt, Jr.[1], C.A. Zukowski[2], and P. Penfield, Jr.[1]

1. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.
2. Department of Electrical Engineering, 1331 S.W. Mudd Bldg., Columbia University, New York, NY 10027.

Abstract

Methods adapted from optimal control are used to calculate simple closed-form upper and lower bounds for the step response of linear systems governed by M-matrices. For high order systems these bounds can be calculated using far less computer time than an "exact" numerical solution requires. The technique has proven to be of real industrial significance in integrated circuit CAD, where it is used for rapid calculation of signal propagation delay through the myriad of interconnect paths in digital MOS IC's.

I. Introduction

Linear systems of differential equations with dynamics governed by M-matrices are commonly used to describe certain diffusion-type systems arising in chemical engineering and biophysics as well as a certain class of electrical RC circuits. Numerically calculating the step response of high order M-matrix systems can consume large amounts of computer time: this is an urgent practical problem in the timing analysis of digital MOS integrated circuits. Driven by this application, a special technique has been developed for rapidly calculating bounds on the step response of a class of linear electrical networks known as RC trees [1], and the results have found widespread practical use, e.g., [2-4]. This paper presents a generalization that applies to a larger class of dynamical systems governed by arbitrary M-matrices.

Many equivalent characterizations of M-matrices can be found in the literature, e.g., [5-11]. We select the following one as a definition because it relates directly to certain manipulations in this paper, and we restrict attention to the nonsingular case.

Def. 1

Let $\underline{A}$ be a nonsingular square matrix of real numbers with $\underline{B} \triangleq \underline{A}^{-1}$. Then $\underline{A}$ is said to be an M-matrix if $a_{jk} \leq 0$, $\forall j \neq k$ and $b_{jk} \geq 0$, $\forall j,k$ [5].

Notation

The symbol $\underline{M}$ will denote any M-matrix, and $\underline{P}$ will denote $\underline{M}^{-1}$. For two vectors $\underline{a}, \underline{b} \in \mathbb{R}^m$, $\underline{a} \geq \underline{b}$ means $a_j \geq b_j$, $j=1,\ldots,m$, and $\underline{a} > \underline{b}$ means $\underline{a} \geq \underline{b}$ and $\underline{a} \neq \underline{b}$. In the dynamical system description (1) below, $\underline{x} \in \mathbb{R}^N$, $\underline{M}$ and $\underline{b}$ are of dimensions compatible with $\underline{x}$, and the subscript i is reserved to denote the distinguished component of $\underline{x}$ selected as the output.

This paper presents a computationally fast method for bounding the zero-state step response of high order single-input single-output linear systems of the form

$$\dot{\underline{x}} = -\underline{M}\,\underline{x} + \underline{b}\,u, \quad \underline{b} > \underline{0} \tag{1}$$

$$y = x_i \ . \tag{2}$$

Simple closed-form expressions in terms of the entries of $\underline{M}$ and $\underline{b}$ will be derived for lower and upper bounds $\underline{x}_i(t)$ and $\bar{x}_i(t)$ such that

$$\underline{x}_i(t) \leq x_i(t) \leq \bar{x}_i(t) \ , \quad \forall t \geq 0 \ , \tag{3}$$

when $u(\cdot)$ is a unit step at $t=0$.

We note that the eigenvalues of any M-matrix have strictly positive real parts [6,8], so (1) is necessarily stable. And it turns out that the step response of (1) is necessarily monotone nondecreasing, although the eigenvalues of $\underline{M}$ need not be real in general.

II. State Equations in the Form Required

The step response bounds will involve the elements of $\underline{P} \triangleq \underline{M}^{-1}$. Multiplying (1) by $\underline{P}$ and rearranging gives

$$\underline{P}\,\underline{b}\,u - \underline{x} = \underline{P}\,\dot{\underline{x}} \ . \tag{4}$$

Note that for a unit step input, $\lim_{t\to\infty} \underline{x}(t) = \underline{P}\,\underline{b}$. Thus we define $\underline{x}_{eq} \triangleq \underline{P}\,\underline{b}$ and obtain the system description

$$\underline{x}_{eq} - \underline{x}(t) = \underline{P}\,\dot{\underline{x}}(t) \ , \quad \forall t \geq 0 \tag{5}$$

$$y = x_i \tag{6}$$

$$\underline{x}(0) = \underline{0} \tag{7}$$

Note that $\underline{x}_{eq} \geq \underline{0}$ as a consequence of (1) and Def. 1. In fact $\underline{x}_{eq} > \underline{0}$ since $\underline{P}$ is nonsingular and $\underline{b} \neq \underline{0}$.

If the state equations are originally given in the form (1), then obtaining $\underline{P}$ requires a time-consuming matrix inversion and obtaining $\underline{x}_{eq}$ requires multiplying a matrix by a vector. But we show in Section VI that these computations can be avoided in a significant class of problems of practical interest, for which the entries of $\underline{P}$ and $\underline{x}_{eq}$ can be obtained directly by inspection of the physical system model.

III. Useful Inequalities for M-Matrix Systems

Fact 1

The zero-state step response of (1) is monotone in time, i.e.,

$$\dot{\underline{x}}(t) \geq \underline{0} \ , \quad \forall t \geq 0 \ . \tag{8}$$

Fact 2

If $\underline{P} \in \mathbb{R}^{N\times N}$ is the inverse of an M-matrix, then

$$P_{ii}P_{kj} \geq P_{ki}P_{ij} \ , \quad \forall i,j,k \in \{1,\ldots,N\} \ , \tag{9}$$

where i, j and k are not necessarily distinct.

Facts 1 and 2 are proved in the Appendix.

Lemma 1

Let $x_i = y$ be the distinguished output variable of (5-7). Then at any instant during the transient the relation between $x_i(t)$ and every other state variable $x_k(t)$ is governed by the inequalities

$$p_{ii}(x_{k_{eq}} - x_k(t)) \geq p_{ki}(x_{i_{eq}} - x_i(t)) \qquad (10)$$

$$p_{ik}(x_{k_{eq}} - x_k(t)) \leq p_{kk}(x_{i_{eq}} - x_i(t)) \quad , \qquad (11)$$

$\forall i,k \in \{1,\ldots, N\}$, $\forall t \geq 0$.

Proof

Expanding the k-th and i-th rows of (5) gives

$$x_{k_{eq}} - x_k(t) = \sum_{j=1}^{N} p_{kj}\dot{x}_j(t) \qquad (12)$$

$$x_{i_{eq}} - x_i(t) = \sum_{j=1}^{N} p_{ij}\dot{x}_j(t) \quad . \qquad (13)$$

Multiply (12) by p_{ii} and (13) by p_{ki} and subtract.

$$p_{ii}(x_{k_{eq}} - x_k(t)) - p_{ki}(x_{i_{eq}} - x_i(t)) =$$

$$\sum_{j=1}^{N} (p_{ii}p_{kj} - p_{ki}p_{ij})\,\dot{x}_j(t) \geq 0, \quad \forall t \geq 0 \quad , \qquad (14)$$

where the last inequality follows from (8) and (9). This proves (10), and (11) follows by interchanging the roles of k and i. ∎

IV. Construction of a Reduced Order Model by Introducing a New State Variable

Assume $x_{i_{eq}} > 0$, since otherwise the step response at the output is identically zero. The goal of this section is to construct a simple second order model of the higher order system (5)-(7) in terms of the variables f_i and g_i, where g_i is a normalized version of the output,

$$g_i(t) \triangleq \frac{x_{i_{eq}} - x_i(t)}{x_{i_{eq}}} \quad , \qquad (15)$$

and f_i is the new state variable

$$f_i(t) \triangleq \int_t^{\infty} g_i(\tau)\, d\tau \quad , \qquad (16)$$

as shown in Fig. 1.

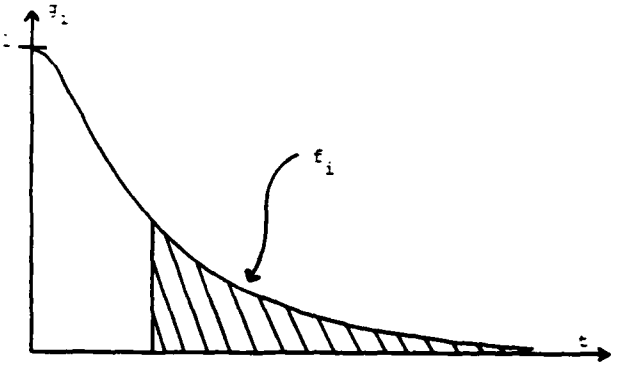


Fig. 1 The variable $f_i(t)$ is the "area to go" under the normalized output $g_i(t)$.

To construct a reduced order model of the system we introduce an artificial input w(t) to represent the unknown waveform $\dot{g}_i(t)$.

Reduced Order Model

Defining $\dot{g}_i \triangleq w$, we assemble (15) and (16) into the reduced order model

$$\begin{bmatrix} \dot{f}_i \\ \dot{g}_i \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_i \\ g_i \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w, \qquad (17)$$

with the constraints and initial conditions below.

Input Constraints

From (8), (15) and the fact that $x_{i_{eq}} > 0$, we conclude that

$$w(t) \leq 0 \quad , \quad \forall t \geq 0 \quad . \qquad (18)$$

Initial Conditions

From (15),

$$g_i(0) = 1 \quad . \qquad (19)$$

From (13) and (16),

$$f_i(0) = \int_0^{\infty} g_i(\tau)\, d\tau = \frac{1}{x_{i_{eq}}} \int_0^{\infty} \sum_{k=1}^{N} p_{ik}\dot{x}_k(\tau)\, d\tau =$$

$$\frac{1}{x_{i_{eq}}} \sum_{k=1}^{N} p_{ik}x_{k_{eq}} \triangleq T_{D_i} \quad . \qquad (20)$$

The time constant T_{D_i} is frequently called the "Elmore delay" [12] in the literature, and the reader can easily verify that it equals the first moment of the normalized impulse response $\dot{x}_i(t)/x_{i_{eq}} = -\dot{g}_i(t)$. For this reason it is used in some applications as an estimate of signal propagation delay through the system [13, 14].

Def. 2

Using the notation in [1], define the additional time constants

$$T_P \triangleq \sum_{k=1}^{N} p_{kk} = \mathrm{tr}\,(P) \qquad (21)$$

$$T_{P_i} \triangleq \frac{1}{p_{ii}} \sum_{k=1}^{N} p_{ik}p_{ki} \qquad (22)$$

Note that $T_{R_i} \leq T_{D_i} \leq T_P$. In fact, these inequalities hold term-by-term for the sums in (20)-(22) defining the time constants, as the reader can verify by evaluating (10) and (11) at t=0.

Lemma 2

The solution to (5)-(7) is such that the state variables f_i and g_i of (17) satisfy the

State Constraints

$$T_{R_i} g_i(t) \leq f_i(t) \leq T_P g_i(t) \ , \quad \forall t \geq 0. \tag{23}$$

Proof

From (13), (15) and (16),

$$f_i(t) = \frac{1}{x_{i_{eq}}} \sum_{k=1}^{N} p_{ik}(x_{k_{eq}} - x_k(t)) \ . \tag{24}$$

Using (10) in (24) yields $f_i \geq T_{R_i} g_i$, and using (11) in (24) yields $f_i \leq T_P g_i$. ■

Equations (17) -(23) define the reduced-order description of the system (5)-(7) and can be analysed with a minimum of computation. See Fig. 2. The cost of this simplification is the introduction of uncertainty, as represented by the inequalities (18) and (23). Thus we obtain only bounds on the true output rather than an exact solution.

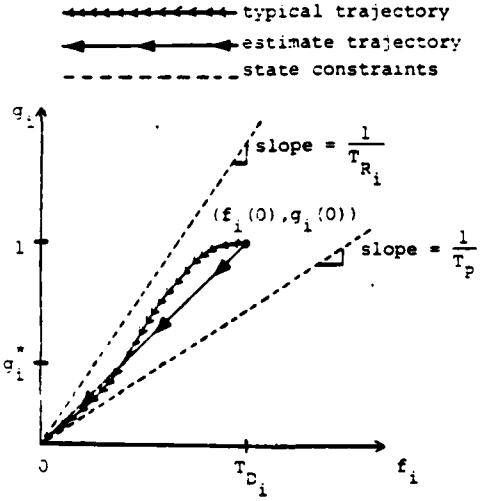


Fig. 2 The step response of a typical high order system of the form (1-2) has the general appearance shown above when plotted in the f_i-g_i plane.

To save computer time we avoid calculating the exact trajectory and use only a) the initial conditions (19) and (20), b) the state constraints (23) as drawn above, and c) the knowledge that the trajectory must move downward because of the input constraint (18) for the reduced order model, and to the left because $f_i = -g_i \leq 0$. The "estimate trajectory" is simple in form and satisfies all these conditions.

V. Optimal Control Method for Determining Step Response Bounds

To calculate bounds on the step response, we first determine the maximum and minimum times at which the state variable $g_i(t)$ in (17) can reach any given "target" value $g_i^* \in (0, 1]$, subject to the constraints (18)-(23). This is a linear minimum- and maximum-time optimal control problem with a fixed initial condition, inequality constraints on the state and input, and terminal condition $g_i = g_i^*$. The optimal trajectories can be determined without recourse to Pontryagin's maximum principle because the time duration t of any path in the f_i-g_i plane can be calculated by rearranging and integrating $df_i/dt = -g_i$ to yield

$$t = \int_{f_{final}}^{f_{initial}} \frac{1}{g_i} df_i \ . \tag{25}$$

Thus the fastest trajectory from the initial state to the "target" interval $(T_{R_i} g_i^* \leq f_i \leq T_P g_i^*, \ g_i = g_i^*)$ is the one for which both the width $\Delta f_i = f_{initial} - f_{final}$ of the region of integration and the integrand $1/g_i$ are minimized. The slowest trajectory is found similarly, and both are displayed in Figs. 3-5. The maximum and minimum times, denoted $t_{max}(g_i^*)$ and $t_{min}(g_i^*)$ because they depend on the "target" value g_i^*, can be calculated directly from the figure using (25). The reader will be spared the algebraic details, which he can easily reproduce if needed: the resulting expressions have appeared in [1] for a slightly special case and can be found in [15] for the general case.

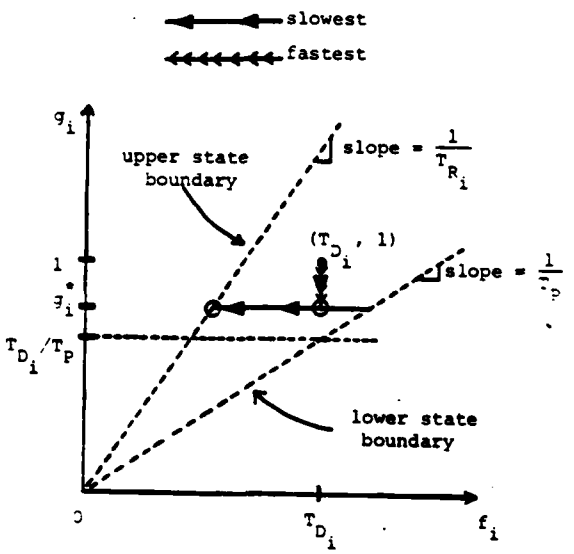


Fig. 3 Form of the maximum- and minimum-time trajectories for the case $T_{D_i}/T_P \leq g_i^* \leq 1$. The common initial state is marked with a dot, the "target" set is the portion of the horizonal line at $g_i = g_i^*$ lying within the state constraints, and the terminal states for the two trajectories are marked with small circles. The minimum-time trajectory drops vertically to the target in zero time. The maximum-time trajectory drops immediately to a point infinitesimally above the target and then moves horizontally to the left at constant velocity.

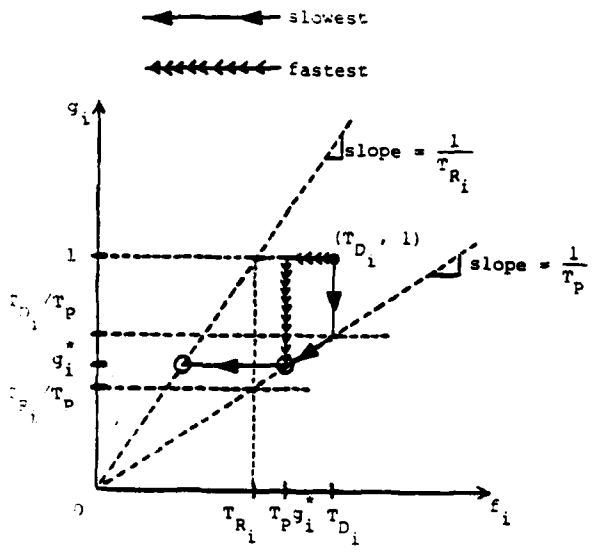


Fig. 4 Form of the maximum- and minimum time trajectories for the case $T_{R_i}/T_P \le g_i^* \le T_{D_i}/T_P$. The fastest trajectory first moves horizontally at constant velocity and then drops vertically to the right hand edge of the target. The slowest trajectory first drops vertically to the lower constraint boundary, proceeds along that boundary until it is infinitesimally above the target, then moves horizontally to the left at constant velocity.

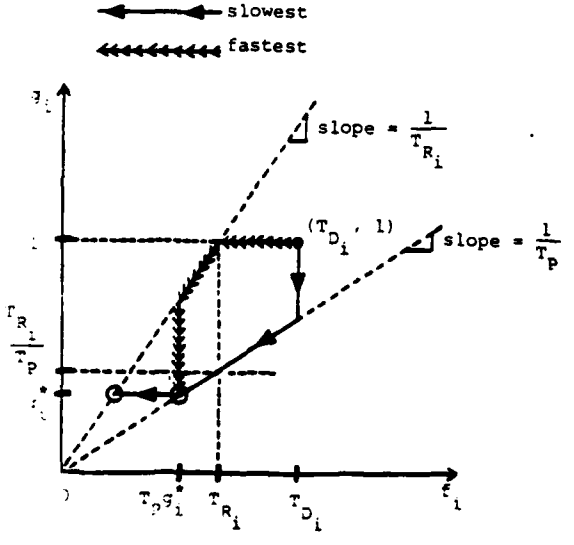


Fig. 5 Form of the extremal trajectories for $0 < g_i^* \le T_{R_i}/T_P$. The slowest trajectory remains similar to Fig. 4, but the fastest trajectory now encounters the upper constraint boundary and proceeds along it before dropping vertically and instantaneously to the right hand edge of the target.

Fortunately, both $t_{min}(g_i^*)$ and $t_{max}(g_i^*)$ can be inverted explicitly to yield upper and lower bounds $\bar{g}_i(t) = t^{-1}_{max}(\cdot)$ and $\underline{g}_i(t) = t^{-1}_{min}(\cdot)$ for $g_i(t)$. These in turn yield lower and upper bounds $\underline{x}_i(t)$ and $\bar{x}_i(t)$ on the step response $x_i(t)$, $\underline{x}_i(t) = x_{i_{eq}}(1 - \bar{g}_i(t)) \le x_i(t) \le \bar{x}_i(t) = x_{i_{eq}}(1 - \underline{g}_i(t))$, $\forall t \ge 0$. The final results are given below and their form is plotted in Fig. 6.

Lower Bound

$$\underline{x}_i(t) = \begin{cases} 0 & , \quad 0 \le t \le T_{D_i} - T_{R_i} \\ x_{i_{eq}}\left(1 - \frac{T_{D_i}}{t + T_{R_i}}\right) & , \quad T_{D_i} - T_{R_i} \le t \le T_P - T_{R_i} \\ x_{i_{eq}}\left(1 - \frac{T_{D_i}}{T_P} e^{\frac{T_P - T_{R_i} - t}{T_P}}\right) & , \quad T_P - T_{R_i} \le t < \infty \end{cases} \tag{26}$$

Upper Bound

$$\bar{x}_i(t) = \begin{cases} x_{i_{eq}}\left(1 + \frac{t - T_{D_i}}{T_P}\right) & , \quad 0 \le t \le T_{D_i} - T_{R_i} \\ x_{i_{eq}}\left(1 - \frac{T_{R_i}}{T_P} e^{\frac{T_{D_i} - T_{R_i} - t}{T_{R_i}}}\right) & , \quad T_{D_i} - T_{R_i} \le t < \infty \end{cases} \tag{27}$$

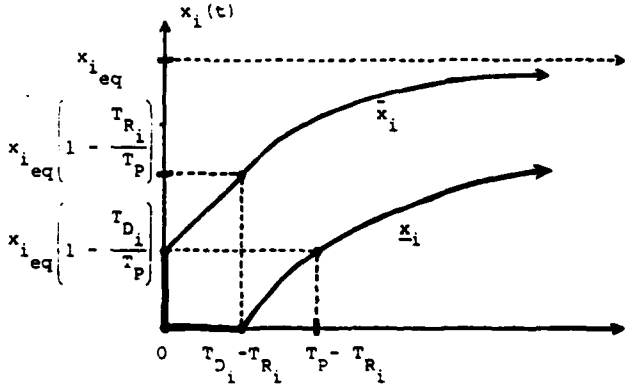


Fig. 6 Form of the step response bounds (26) and (27). The distance between them is exaggerated for clarity.

Single Time Constant Estimate for the Step Response

The simple time function

$$\hat{x}_i(t) \triangleq x_{i_{eq}}(1 - e^{-t/T_{D_i}}), \quad \forall t \ge 0, \tag{28}$$

generates a straight line from the initial condition to the origin when plotted in the f_i-g_i plane, as shown in Fig. 2. When plotted against t, it shares the following features with the exact step response $x_i(t)$: both have value zero at t=0 and rise smoothly and monotonically to

$x_{i_{eq}}$ as $t \to \infty$, and the first moment of both $\dot{x}_i(t)/x_{i_{eq}}$ and $\dot{\hat{x}}_i(t)/x_{i_{eq}}$ is T_{D_i}. For this reason $\hat{x}_i(\cdot)$ is frequently used as an estimate of the exact step response waveform $x_i(\cdot)$ [13,14]. Note from Fig. 2 that $\hat{x}_i(\cdot)$ is always a feasible (but not generally optimal) solution to the constrained optimal control problem and hence must always lie between the step response bounds (26) and (27).

The estimate $\hat{x}_i(\cdot)$ is an attempt to approximate the step response of a high order system by that of one with a single time constant: the bounds measure the worst case error resulting from that approximation. Whenever $T_P - T_{R_i} \ll T_{D_i}$ the wedge-shaped feasible region in Figs. 2-5 becomes narrow, the maximum- and minimum-time trajectories lie close together, and the bounds become very tight: this is commonly the case in the application below. The interested reader may wish to verify the following relation between the time constants and the tightness of the bounds:

$$\sup_{t \geq 0} [\bar{x}_i(t) - \underline{x}_i(t)] \leq x_{i_{eq}} \left[\frac{T_P}{T_{R_i}} - 1 \right] . \qquad (29)$$

VI. Application to VLSI CAD

This paper expands and generalizes previous research [1] focussed on the task of estimating signal propagation delay in branching interconnect lines on MOS VLSI chips. The earlier theory relied on special features of interconnect networks not shared by the general M-matrix systems discussed here. This section is confined to a brief description of the application: full details are given in [1].

A reasonable electrical model for interconnect paths on chips is a linear, nonuniform "RC tree" network consisting of a tree of linear resistors driven by an ideal voltage source and shunted by capacitors to ground, as illustrated in Fig. 7. As many as 100,000 of these interconnect nets can be found on a single chip, and they come in a great variety of sizes and configurations. The propagation delay from the source to any node functioning as an output is determined from the step response at that node in conjunction with knowledge of the switching voltage threshold of the logic gate driven by that output. The bounds have been found useful in practice [2-4] because they save computer time: exact numerical calculation of the step response of such a vast number of systems of moderately high order is out of the question.

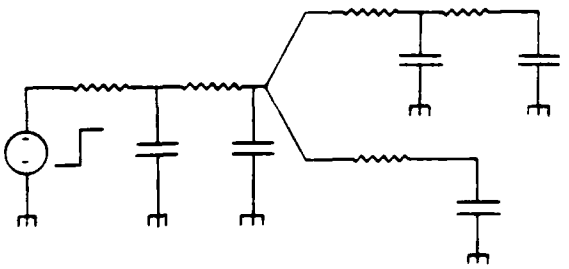


Fig. 7 Lumped, linear RC tree networks are useful models for branched transmission lines on MOS chips: their dynamics are described by M-matrices.

It is not difficult to show [16] that the dynamics of such networks are gover - by equations of the form (1), (2): of greater rest here is the fact that the entries of $P = M^{-1}$ and of x_{eq} in (5) can be found by inspection. Using the capacitor voltages as state variables, we see that no current flows in the resistors at equilibrium, so the equilibrium voltages are all 1 V. for a unit step input. And it is easy to show [1], [16] that the state equations for the step response of any N-capacitor RC tree can be written in the form

$$(\underline{1} - \underline{v}) = [\underline{RC}] \dot{\underline{v}} , \qquad (30)$$

where $\underline{v} \in \mathbb{R}^N$ is the vector of capacitor voltages, $\underline{1} \in \mathbb{R}^N$ is a vector of 1's, $\underline{C} \in \mathbb{R}^{NxN}$ is a diagonal matrix of the capacitor values, and $\underline{R} \in \mathbb{R}^{NxN}$ is obtained by setting R_{jk} equal to the sum of the resistances along the path obtained by intersecting the route from C_j to the source with the route from C_k to the source. Thus $\underline{P} = [\underline{RC}]$ is obtained without matrix inversion.

VII. Additional Results

The alert reader may have noticed that the extremal trajectories in Figs. 3-5 contain unrealistic vertical segments representing discontinuous jumps in g_i. A practical method of tightening the bounds by incorporating finite slew rate limits was briefly described in [17] along with a second bound-tightening technique applicable only to RC trees.

A way to apply the results in this paper to an important class of nonlinear RC circuits is given in [18].

Acknowledgement

It is a pleasure to acknowledge helpful conversations with Professor Qingjian Yu of the East China Institute of Technology, Nanjing, People's Republic of China, Professor Han Ngee-Tan of the Nanyang Technological University in Singapore, and Professor Violet Haas of Purdue University. And thanks to Peter O'Brien and Keith Nabors for assistance with earlier versions of this manuscript. This research was supported by the National Science Foundation under Grant No. ECS83-10941 and the Defense Advanced Research Projects Agency under Contract No. N00014-80-C-0622.

Appendix

Proof of Fact 1

Consider (1) with a unit step input at t=0. Differentiating with respect to t yields the equation governing the evolution of $\dot{x}$,

$$\frac{d}{dt} \dot{\underline{x}} = -\underline{M}\dot{\underline{x}} , \quad \forall t > 0 , \qquad (A.1)$$

$$\dot{\underline{x}}(0^+) = \underline{b} > \underline{0} . \qquad (A.2)$$

Because the off-diagonal elements of $-\underline{M}$ are nonnegative, the closed first orthant of $\mathbb{R}^N$, usually denoted $\mathbb{R}^N_+$, is positive invariant under the flow of (A.1) [19]. Thus $\dot{\underline{x}}(0^+) \geq \underline{0} \Rightarrow \dot{\underline{x}}(t) \geq \underline{0}$, $\forall t \geq 0$. ■

Proof of Fact 2

If i=j or i=k, (9) is trivial. If j=k, (9) becomes $P_{ii}P_{kk} \geq P_{ki}P_{ik}$, which is true because the determinant of any principal submatrix of an inverse M-matrix is known to be positive [11, Cor. 1, p. 198].

If i, j and k are all distinct, consider the 3x3 principal submatrix of $\underline{P}$ consisting of elements that lie both in row i, j or k and in column i, j or k, and denote it $\hat{P}$. Since any principal submatrix of an inverse M-matrix is known to be an inverse M-matrix [5, p. 329], $\hat{P}$ is "inverse M" and hence $(\hat{P}^{-1})_{kj} \leq 0$. Taking the inverse of $\hat{P}$ using cofactors, we have $0 \geq (\hat{P}^{-1})_{kj} = \text{cof } p_{jk}/\det \hat{P}$, and $\det \hat{P} > 0$ [11, p.198]. Thus $\text{cof } p_{jk} = p_{ki}p_{ij} - p_{ii}p_{kj} < 0$, proving (9). ■

References

[1] Rubinstein, J., P. Penfield, Jr., and M.A. Horowitz, "Signal Delay in RC Tree Networks, IEEE Trans. Computer-Aided Design, vol. CAD-2, no. 3, pp. 202-211, July 1983.

[2] Putatunda, R., "Autodelay: A Second-Generation Automatic Delay Calculation Program for LSI/VLSI Chips," Proc. IEEE Int. Conf. on Computer-Aided Design, Santa Clara, CA, November, 1984, pp. 188-190.

[3] Jouppi, N., "Timing Analysis for NMOS VLSI," ACM IEEE 20th Design Autom. Conf. Proc., June 1983, pp. 411-418.

[4] Tamura, E., K. Ogawa, and T. Nakano, "Path Delay Analysis for Hierarchical Building Block Layout System," ACM IEEE 20th Design Autom. Conf. Proc., June 1983, pp. 403-410.

[5] Markham, T.L., "Nonnegative Matrices Whose Inverses are M-Matrices," Proc. Amer. Math. Soc., vol. 36, no. 2, Dec. 1972, pp. 326-330.

[6] Araki, M., "M-Matrices (Matrices with Nonpositive Off-Diagonal Elements and Positive Principal Minors)," Publ. 74/19, Dept. of Computing and Control, Imperial College, London, March 1974.

[7] Willems, J.C., "Lyapunov Functions for Diagonally Dominant Systems," Automatica, vol. 12, 1976, pp. 519-523.

[8] Plemmons, R.J., "M-Matrix Characterizations. I-Nonsingular M-Matrices," Linear Algebra and Its Applications, vol. 18, 1977, pp. 175-188.

[9] Willoughby, R.A., "The Inverse M-Matrix Problem," Linear Algebra and Its Applications, vol. 18, 1977, pp. 75-94.

[10] Araki, M., "Stability of Large-Scale Nonlinear Systems-Quadratic Order Theory of Composite System Method Using M-Matrices," IEEE Trans. Automatic Control, vol. AC-23, no. 2, April 1978, pp. 129-142.

[11] Johnson, C.R., "Inverse M-Matrices," Linear Algebra and Its Applications, vol. 47, 1982, pp. 195-216.

[12] Elmore, W.C., "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers," Jour. Appl. Physics, vol. 19, pp. 55-63, January 1948.

[13] Lin, T.-M., and C.A. Mead, "Signal Delay in General RC Networks," IEEE Trans. Computer-Aided Design, vol. CAD-3, no. 4, October 1984, pp. 331-349.

[14] Ragunathan, A. and C. Thompson, "Signal Delay in RC Trees with Charge Sharing or Leakage," Report No. UCB/CSD 85/243, Computer Science Division, Univ. of California, Berkeley, CA 94720, June 1985.

[15] O'Brien, P. and J.L. Wyatt, Jr., "Signal Delay in Leaky RC Mesh Models for Bipolar Interconnect," VLSI Memo, Sept. 1985. All VLSI Memos are available from the Microsystems Program Office, Room 39-321, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139.

[16] Wyatt, J.L., Jr., "Signal Delay in RC Mesh Networks," IEEE Trans. Circuits and Systems, vol. CAS-32, no. 5, May 1985, pp. 507-510.

[17] Yu, Q., J.L. Wyatt, Jr., C. Zukowski, H.N. Tan and P. O'Brien, "Improved Bounds on Signal Delay in Linear RC Models for MOS Interconnect," Proc. 1985 IEEE Int. Symp. on Circuits and Systems, Kyoto, Japan, June 1985, pp. 903-906.

[18] Wyatt, J.L., Jr., "Monotone Sensitivity of Nonlinear Nonuniform RC Transmission Lines, with Application to Timing Analysis of Digital MOS Integrated Circuits," IEEE Trans. Circuits and Systems, vol. CAS-32, no. 1, Jan. 1985, pp. 28-33.

[19] Bellman, R., Introduction to Matrix Analysis McGraw-Hill, New York, 1970, p. 176.

END

FILMED

1-86

DTIC